First International Workshop on Video Processing for Security (VP4S-06), June 7-9, Quebec City, Canada, 2006.

3D Face Reconstruction from Stereo Video

Unsang Park and Anil K. Jain
Computer Science and Engineering
Michigan State University
{parkunsa, jain}@cse.msu.edu

Abstract

Face processing in video is receiving substantial attention due to its importance in many security-related applications. A video provides rich information about a face (multiple frames and temporal coherence) that can be utilized in conjunction with 3D face models, if available, to establish a subject's identity. We propose a 3D face modeling method that reconstructs a user-specific model derived from a generic 3D face model and two video frames of the user. The user-specific 3D face model can be enrolled into the 3D face database at the enrollment stage to be used in later identification process. The reconstruction process can also be used for the probe data in recognition stage, where the reconstructed 3D face model using probe face is used to generate an optimal view and lighting for the recognition process. The advantage of utilizing reconstructed 3D face model is demonstrated by conducting face recognition experiments for 15 probe subjects against a gallery database containing 100 subjects.

1. Introduction

With increasing demands for higher security in a number of critical applications, developing a robust method of utilizing face as a biometric has emerged as an important research issue. Recent advances in 3D sensing technology and a better understanding of 3D geometry have shifted the focus of face recognition from 2D to 3D and 2D/3D hybrid domain. There are a number of advantages in using 3D face models. Three-dimensional model based face recognition is robust against pose and lighting variations. The identification can be performed between two (2.5D) range (depth) images or between a 2D image and the 3D face model [1]. Even though 3D face modeling with the assistance of 3D sensing devices (range sensors) has proved its effectiveness for 3D face recognition, the limited applicability and high cost of these sensors underscore the importance of 3D face reconstruction from 2D images or video streams.

There have been many studies on 3D face recognition using reconstructed 3D models from a set of 2D images [2], [3], from multiple video frames [4], [5] or other methods [6]. The reconstructed 3D model is used to generate the 2D projection images that are matched with (2D) probe images [1]. Alternatively, the reconstructed 3D model can be used to generate a frontal view of the probe image with arbitrary pose and lighting; the recognition is then performed with the synthesized frontal faces. A combination of 2D and 3D face recognition systems is also regarded as a promising method [7]. A 2D and 3D hybrid matching method described in [7] used the 2D projection images of the 3D model to construct the LDA subspace for 2D face recognition.

It has been shown that utilizing multiple face images per subject with varying pose and illumination can effectively handle the pose and lighting variations observed in medium quality video [9]. Since multiple face images with various pose and lighting can be synthesized from a 3D face model, acquiring an accurate 3D face model is a prerequisite. In [9], 3D face models of all the subjects were acquired during enrollment using a range sensor. However, some subjects of interest may not be available for enrollment using the range sensor, so alternate methods of obtaining 3D face model need to be explored.

3D face reconstruction problem for the purpose of face recognition consists of the following main components: i) face and facial feature point detection, ii) 2D to 3D reconstruction and iii) utilizing 3D face model for recognition. In this paper, we focus on parts ii) and iii) and allow manual interventions for part i) for better control and evaluation in the face reconstruction and recognition process.

We propose a fast and realistic 3D face reconstruction method by fitting a semi-dense generic

face model to specific subjects through a set of reconstructed facial landmarks observed from two-view video. The reconstructed facial landmarks are used as a set of control points for the Thin Plate Spline (TPS) [10] fitting process. Our method of 3D face reconstruction to assist face recognition has some advantages compared to other approaches: i) higher quality of reconstructed 3D face than [4], ii) faster processing time than [3] and iii) not requiring user's range data as in [9], [6]. We have evaluated the proposed 3D face modeling method on a probe database of 15 subjects with gallery database containing 100 subjects, which is larger than the gallery sizes in [1] and [2].

This paper is composed as follows. Section 2 introduces 3D face model reconstruction process. Section 3 introduces face recognition process. Section 4 describes the experimental results and section 5 concludes the paper.

2. 3D Face Reconstruction

Our 3D face modeling process starts with reconstructing sparse set of facial landmark points from stereo video. The reconstructed 3D facial landmarks are fed to the generic face model and the generic model undergoes non-linear transformation process based on TPS. Finally, the adapted 3D face model is mapped with the texture data available from the video to generate realistic user-specific 3D face models. The overall process of 3D face reconstruction is depicted in Fig. 1.

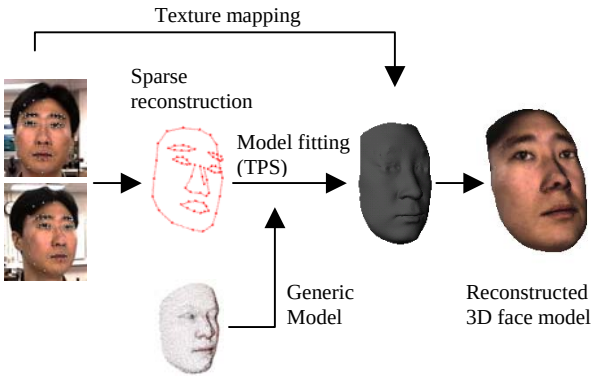


Figure 1. System diagram of 3D face model reconstruction.

2.1. Data Collection

The two cameras used to obtain a pair of stereo images are calibrated with an open source package. [11] Videos of 15 subjects are recorded from two cameras and a pair of images is selected from the video stream for the 3D face reconstruction experiment. The pair of images is selected to have one of the two images show near frontal pose in order to ensure the texture image of frontal view is available.

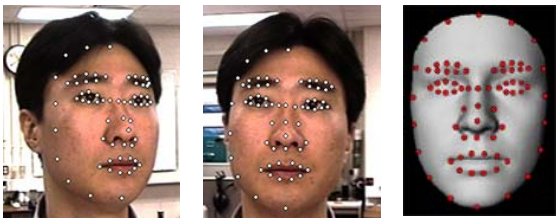


Figure 2: Facial feature points overlaid on the (a) face image from left camera, (b) face image from right camera, and (c) generic model.

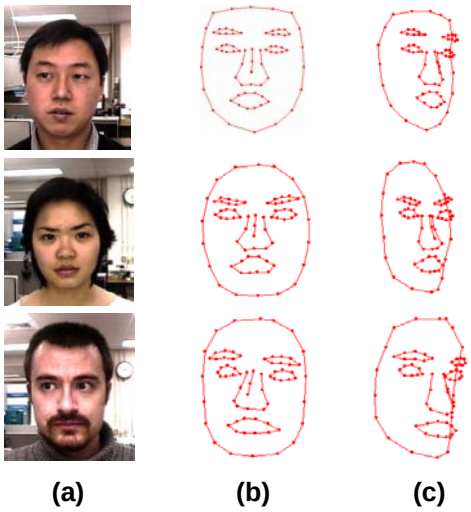


Figure 3. Sparse facial feature point reconstruction from stereography. (a) raw image, reconstructed 3D geometry at (b) frontal pose, and (c) left pose.

2.2. Sparse Point Set Reconstruction

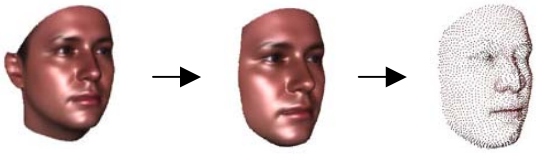
Our reconstruction scheme for the facial landmarks is based on a set of known corresponding point pairs. Facial landmark points from multiple face images can be obtained automatically by adapting one of the available facial feature point localization approaches [12], [13]. However, we currently use manually selected facial feature points and focus on the reconstruction process and its effect on face recognition accuracy.

We select up to 64 feature points in each pair of 2D images, where 54 out of 64 points capture the semantic structure of the internal facial features and the remaining 10 points give the face boundary as shown in Figs. 2 (a) and (b). With the viewpoint changes, some parts of the face boundary are occluded and the corresponding point pairs are not available. In the reconstruction process, only half of the face boundary is recovered and the other half is interpolated based on the facial symmetry to overcome the occlusion of facial boundary.

From the camera projection matrices and a set of point correspondences, 3D shape of facial feature points is obtained from a closed form equation as

$$\begin{bmatrix} P_3 x - P_1 \\ P_3 y - P_2 \\ P_3' x' - P_1' \\ P_3' y' - P_2' \end{bmatrix} X = 0, \qquad (1)$$

where P_i represents the i^{th} row vector of the calibration matrix, x and y represent the pixel coordinates in the image and X represents the 3D coordinates of each corresponding point pair. Some example images of reconstructed sparse facial feature points are shown in Fig. 3.



(a) (b) (c)

Figure 4. Generic model generation; (a) original mean face of the morphable model with ~70,000 vertices, (b) trimmed model with ~40,000 vertices and (c) a model when the number of vertices is reduced to 5,000.

2.3. Generic 3D Face Model

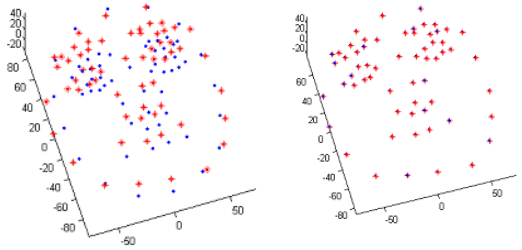
The generic model can be chosen from any 3D face model with a reasonable number of vertices. We used the average 3D face model of the Morphable face model [14] in our experiment. By using the average face model, it is expected that the overall deformation in the fitting process will be minimal. Fig. 4 shows the process of generating generic model from the mean shape of the morphable model. The full 3D model is trimmed along the face boundary (e.g., to delete ears and neck) and then the number of vertices is reduced to about 5,000. The ears and neck are removed because their reconstruction does not lead to any improvement in face recognition. The reduced number of vertices makes the fitting process faster without losing significant amount of information. From the generic model, 72 control points are selected for the model fitting process. The 72 control points are shown overlaid on the generic model in Fig. 2 (c).

2.4. Coarse Alignment

The reconstructed facial feature points are coarsely aligned with the generic model before the fine adaptation process. Let X_1={x∈left eyebrow or x∈left eye}, X_2={x∈right eyebrow or x∈right eye} and X_3={x∈mouth} be three semantic sets of points belonging to the set of control points in generic model. The corresponding three points sets are defined as X_1', X_2' and X_3' in the target model. The coarse alignment minimizes

$$D = \frac{1}{3}\sum_{i=1}^{3} \left\| mean(X_i) - mean(X_i') \right\|. \qquad (2)$$

Given two sets of three points with known correspondences, the average distance D can be minimized by using the approach in [16].



(a) (b)

Figure 5. Fitting Control points using TPS. (a) Point sets before the fitting, and (b) after fitting. Blue (cross hair) points belong to the generic model and red (dot) points belong to the target model.

2.5. Generic Model Fitting Using TPS

Our generic model fitting process relies on learning the deformation between a set of control points in generic model and the target model. Let X = {x_i | i=1, 2, …,n} be the control points on the generic model and Y be the control points on the target model. The deformation between these two point sets can be obtained by the mapping function F(u):

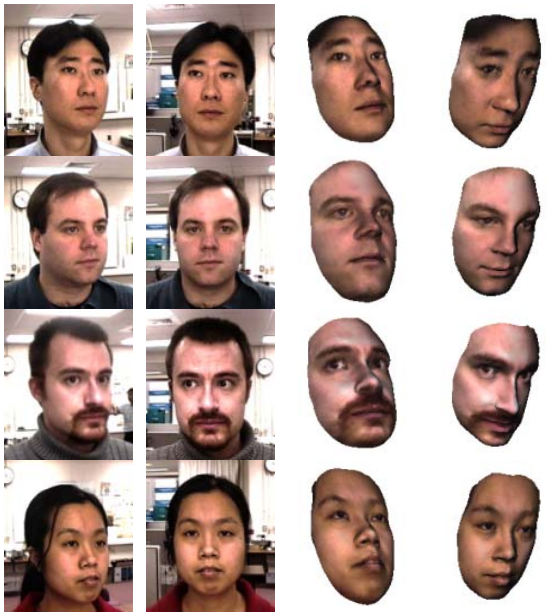
$$F(u) = c + Au + W^T s(u), \quad (3)$$

where c accounts for translation, A for rotation, W^T for the non-linear deformation and $s(u)$ is expressed as:

$$s(u) = (\sigma(u-u_1), \sigma(u-u_2), \ldots, \sigma(u-u_n))^T \quad (4)$$

and $\sigma(u)$ is defined as:

$$\sigma(u) = \begin{cases} \|u\|^2 \log(\|u\|) & \|u\| > 0 \\ 0 & \|u\| = 0. \end{cases} \quad (5)$$



(a) (b) (c) (d)

Figure 6. Reconstructed 3D face models based on generic model of Fig. 4. 2D face images from (a) left camera and (b) right camera. Reconstructed 3D face models at about 30° (c) upper to left, and (d) down to right pose.

With two additional constraints, $1_n^T W=0$ and $U^T W=0$, parameters in Eq. (3) can be solved by:

$$\begin{bmatrix} H & 1_n & U \\ 1_n^T & 0 & 0 \\ U^T & 0 & 0 \end{bmatrix} \begin{bmatrix} W \\ c^T \\ a^T \end{bmatrix} = \begin{bmatrix} V \\ 0 \\ 0 \end{bmatrix}. \quad (6)$$

This procedure is well known as the Thin-Plate Spline (TPS) [10] method. Once the mapping function F(u) is obtained, all the other vertices in the generic model are mapped by F(u) and the final reconstructed 3D face model is obtained. Fig. 5 shows the deformation of control points through TPS.

2.6. Texture mapping

Texture is very important information in face recognition. Therefore, the user-specific 3D face model needs to be augmented with a proper texture to appear realistic. The reconstructed 3D face model is aligned with one of the texture image that is used to reconstruct the 3D shape of facial feature points. The same set of control points as shown in Fig. 2 is used for the alignment process and all vertices in the generic model are projected onto the texture image and assigned the corresponding color values. Some example 3D face reconstruction results are shown in Fig. 6.

(a) (b) (c)

Figure 7. A 3D face model of a randomly selected subject used for the reconstruction at (a) right pose, (b) frontal and (c) left pose.

Figure 8. Reconstructed 3D face models of the same subjects shown in Fig. 6. Reconstruction is performed by using the arbitrarily selected 3D face model as shown in Fig 7 as opposed to generic model shown in fig. 4.

2.6. Generic versus specific model

The effect of the initial 3D face model on the proposed reconstruction is investigated by comparing the reconstruction result of using generic model (Fig. 6) with that of using a specific user's model (Fig. 8). A

3D face model is randomly selected from our gallery database and used as the generic model for the reconstruction. The specific subject's model and reconstruction results are shown in Fig. 7 and Fig. 8, respectively. The reconstructed face models using specific user's 3D model show close resemblance to those using generic model with a slight degradation in the quality. We performed the remaining experiments based on generic 3D face model.

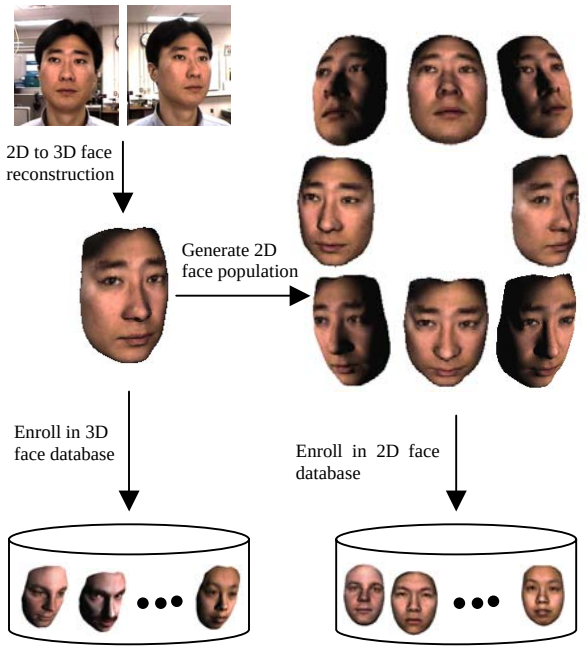


Figure 9. System diagram for the enrollment process using 3D face reconstruction from stereo images.

3. Face Recognition

3.1. Overall System

The proposed 3D face reconstruction system can be used both in enrollment and identification stages in either 2D or 3D domain. Some of the components that can be improved using the 3D reconstruction scheme are i) enrollment of 3D face model using a 2D camera and ii) enrollment of 2D face images with various pose and lighting conditions. The system diagram for the enrollment process using 3D face reconstruction from video is shown in Fig. 9. In the identification stage, the matching process can be performed both in 2D and 3D domains.

3.2. Face Recognition in 2D domain

Fifteen video files were recorded from fifteen subjects under three different lighting conditions at various poses with yaw and pitch motion. A subset of pose and lighting variations is captured from the video for the design and evaluation purposes for the face recognition system. The selected variations are about 20 degrees to the right, left, up, and down under 4 different lighting conditions. The lighting conditions we employed are normal, dark, and light source at 45 and 90 degrees from front to right. The sample frames from the video of one subject are provided in Figure 10.

Gallery data is prepared in two sets: one from the real 3D face models and the other from the adapted generic face models combined with real 3D face models. The "real" face model represents the 3D model constructed from the range sensor. The range data of each subject is captured at frontal, right pose and left pose. These three range images are stitched together by an interactive tool to generate the real 3D model. First gallery data consists of 100 real 3D face models including 15 probe subjects. Second gallery data consists of reconstructed 3D face models of 15 probe subjects and additional real 3D face models of 85 subjects.

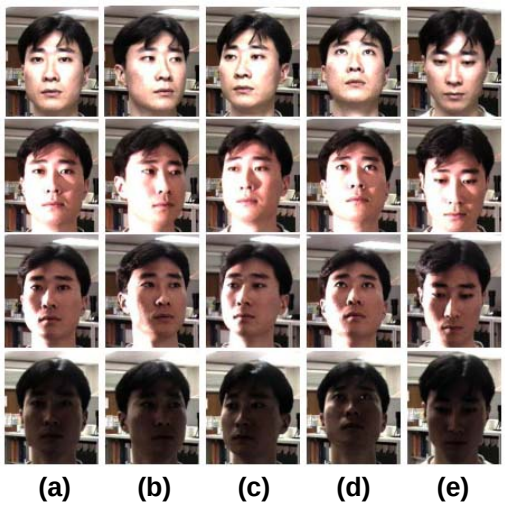


Figure 10. Pose and lighting variations observed in the video of one subject. (a) Frontal pose, (b) Left pose, (c) Right pose, (d) Up pose and (e) down pose. First row shows normal lighting, second and third row shows images with a light source at 45 and 90 degrees from frontal to right and the last row shows images under dark lighting.

Accurate eye locations are needed both from probe and gallery images for proper alignment in the face recognition process. Currently, we use manually selected eye locations to minimize the effect of feature extraction errors on the face recognition performance.

Let $\chi = \{\chi_i, i=1,2,\ldots,N_\chi\}$ be the set of 3D models enrolled in the gallery. Given a 2D probe image *p*, the identity is decided by

$$ID = \arg \max_{i=1,2,\ldots,N_\chi} s(p, \chi_i), \quad (7)$$

where s(·) represents the matching score measured by FaceIt® SDK [15] or FaceVAC® SDK [17]. The matching scores obtained from the two face recognition SDKs are combined using the sum method after min-max normalization [8]. Having a set of 2D projection images $\lambda_i = \{\lambda_{ij}, j=1, 2,\ldots,N_{\lambda i}\}$ for each model χ_i, the identity can be equivalently decided by

$$ID = \arg \max_{i=1,2,\ldots,N_\chi} \left\{ s(p, \lambda_{i,j}), \exists j, j = 1,2,\ldots,N_{\lambda i} \right\} \quad (8)$$

We generate 20 projection images from each of the 100 3D models to build the gallery. Therefore, N_χ is equal to 100 and $N_{\lambda i}$ is equal to 20.

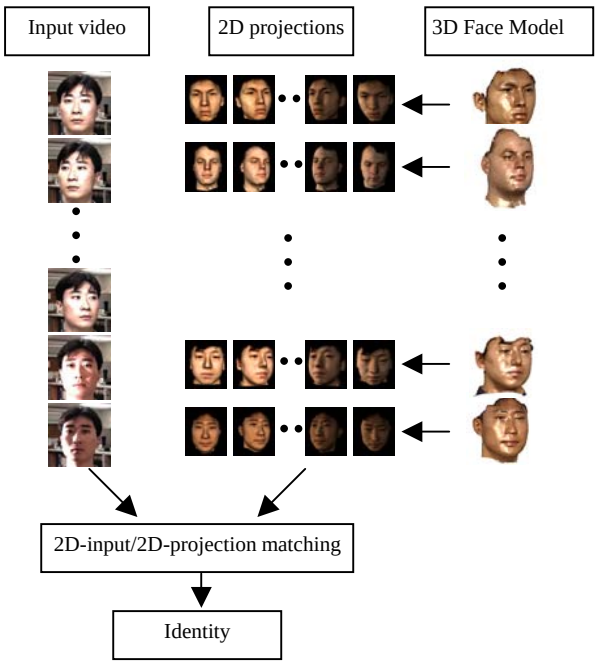


Figure 11. Face recognition system using video input and 3D model gallery.

To utilize the information contained in the input video, majority-voting rule is used to fuse the recognition results from multiple frames. Among the various fusion rules, it has been shown that the simple majority voting achieves good performance [9]. In majority voting, the maximum matching scores are computed for a set of probe images $\gamma = \{\gamma_k, k=1,2,\ldots,N_\gamma\}$ as in Eq. (8). Then, the identity is decided from the majority *ID* in the set of probe images as

$$ID_{mv} = majority\{ID_k, k = 1,2,\ldots,N_\gamma\}. \quad (9)$$

The face recognition process using video and 3D face gallery is depicted in Fig. 11.

3.3. Face Recognition in 3D domain

Given the reconstructed 3D face model from video, face recognition can be performed in 3D domain by matching the reconstructed 3D face model at identification time against the one reconstructed at enrollment time. The matching also can be performed between a reconstructed face model from video against an enrolled 3D face model from other 3D sensing device (e.g., 3D laser scanner).

4. Experimental Results

Experimental results are provided in two stages: one is the 2D to 3D face reconstruction and the other is the face recognition by utilizing the reconstructed face models. 3D face reconstruction results were already provided in Figs. 6 and 8. The proposed reconstruction process takes about 2.95 seconds on average on a Pentium 4 3.2GHz PC besides the time for the manual feature point selection.

Fig. 12 shows the rank-N accuracies of face identification test with and without pose and lighting variations. The baseline 2D face matcher (score-sum of FaceIt and FaceVAC) performs very well (with rank-1 accuracy of 93%) when both probe and gallery faces are frontal under normal lighting condition. However, the performance drops drastically to rank-1 accuracy of 38% when the probe images have pose and lighting variations. This is the main problem that is being addressed here.

Fig. 13 shows the performance with using multiple frames as probe and utilizing the real 3D face model to populate the pose and lighting variations in the gallery. Fig. 14 shows the same experimental results by using reconstructed 3D face models for the 15 probe subjects instead of the real 3D models. The performances

shown in Figs. 13 and 14 with single frame are better compared to Fig. 12 (all pose and lighting with frontal gallery) by the effect of pose and lighting variations in gallery data. The performance with reconstructed face models (rank-1 accuracy of 78% when using 30 frames) is lower than the performance with real 3D face models (rank-1 accuracy of 83% when using 30 frames). Example images of the real 3D model and reconstructed 3D model are shown in Fig. 16. Based on the higher face recognition performance from real face models than the reconstructed models, we believe that the noise involved in the reconstruction process is responsible for this degradation in the face recognition performance. Some example matching results are shown in Fig. 15, where the identification fails when the gallery contains only frontal images, but succeeds with gallery containing images with various pose and lighting variations.

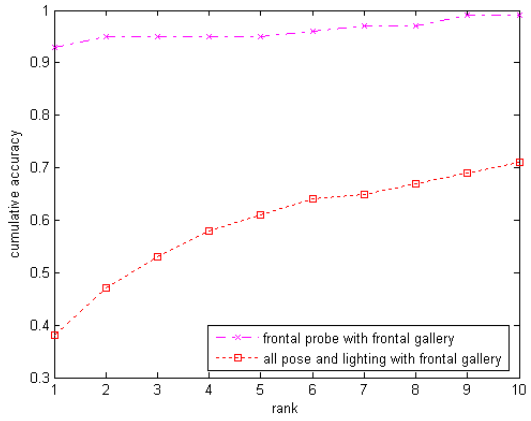


Figure 12. Rank-N accuracy with and without pose and lighting variations in the probe data.

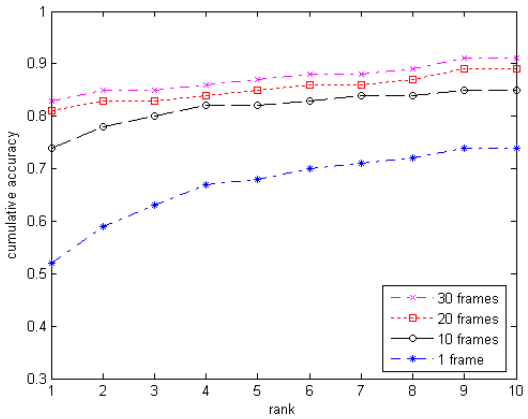


Figure 13. Rank-N accuracy by fusing different numbers of frames and real 3D face models.

5. Conclusions and Future Work

We have proposed a 3D face reconstruction method and showed its effectiveness in face recognition in video. The reconstruction process uses semi-dense generic model, and hence provides better quality of reconstructed model and fast processing time.

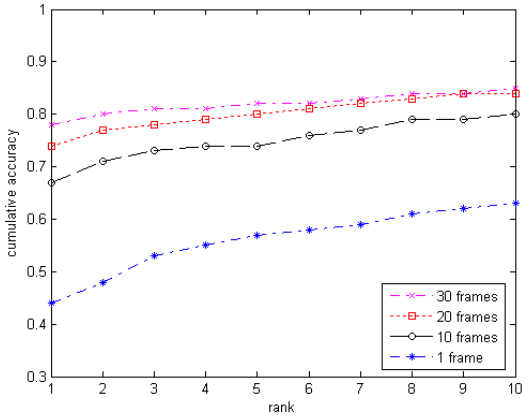


Figure 14. Rank-N accuracy by fusing different numbers of frames and reconstructed 3D face models.

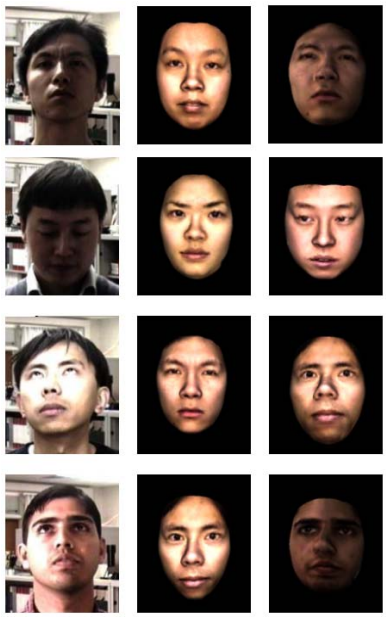
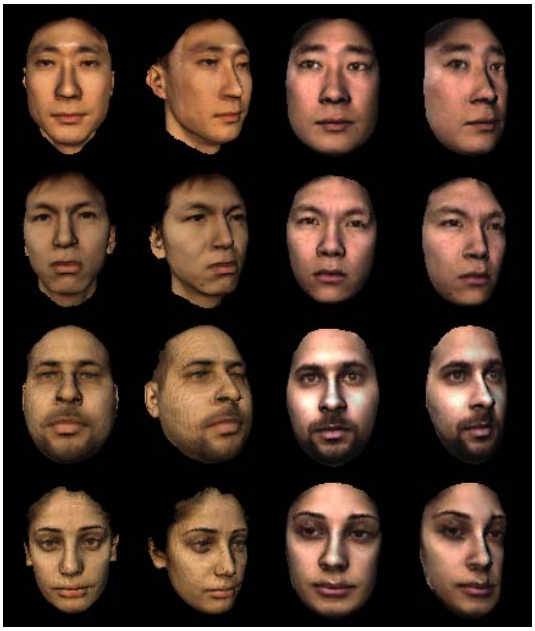


Figure 15. Examples of matching results with reconstructed 3D face models. (a) Raw image from video, (b) Incorrect matches with the gallery having only frontal faces, and (c) Correct matches with the gallery containing pose and lighting variations.

The reconstructed model is able to populate the gallery with 2D images with pose and lighting variations, resulting in improved recognition accuracy.

Future work will include developing a more robust and accurate reconstruction method, fully automating the reconstruction and recognition system from the video input with fewer constraints, and improving the system to work under arbitrary pose and lighting situations.



(a) (b) (c) (d)

Figure 16. Reconstructed vs. real 3D face models. (a) and (b) Example images of real 3D models and (c) and (d) reconstructed 3D models.

References

[1] Mun Wai Lee and Surendra Ranganath, "Pose-invariant face recognition using a 3D deformable model," Pattern Recognition, 2003, 36, pp. 1835-1846.

[2] A. Georghiades, P.N. Belhumeur and D. Kriegman, "From few to many: Generative models for recognition under variable pose and illumination," Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000, pp. 277-284.

[3] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3D morphable model," IEEE Transactions on PAMI, 2003, 25(9), pp. 1063-1074.

[4] A. Roy Chowdhury and R. Chellappa, "Face reconstruction from video using uncertainty analysis and a generic model," Computer Vision and Image Understanding, 2003, 91(1-2), pp. 188-213.

[5] M. Brand, "Morphable 3D models from video," in Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, II, pp. 456-463.

[6] R.-L. Hsu and A. K. Jain, "Face modeling for recognition", in Proc. International Conference on Image Processing (ICIP), Greece, 2001, pp. 693-696.

[7] X. Lu, A. K. Jain and D. Colbry, "Matching 2.5D Face Scans to 3D Models," IEEE Transactions on PAMI, 2006, 28(1), pp. 31-43.

[8] A. K. Jain, K. Nandakumar and A. Ross, "Score Normalization in Multimodal Biometric Systems", Pattern Recognition, December 2005, 38(12), pp. 2270-2285.

[9] U. Park, H. Chen and A.K. Jain, "3D Model-assisted Face Recognition in Video," Proc. of 2nd Workshop on Face Processing in Video, Victoria, British Columbia, Canada, 2005, pp. 322-329.

[10] F.L. Bookstein, "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations," IEEE Transactions on PAMI, 1989, 11(6), pp. 567-585.

[11] Camera calibration toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/

[12] S. C. Yan, M. J. Li, H. J. Zhang and Q. S. Cheng. "Ranking Prior Likelihood Distributions for Bayesian Shape Localization Framework," in Proc. Intl. Conf. on Computer Vision, France, Nice, 2003, 2, pp. 51-58.

[13] D. Cristinacce, T.F. Cootes and I. Scott, "A Multistage Approach to Facial Feature Detection" Proc. British Machine Vision Conference, 2004, 1, pp. 277-286.

[14] V. Blanz and T. Vetter. "A morphable model for synthesis of 3D faces." Computer Graphics Proceedings SIGGRAPH, Los Angeles, 1999, pp. 187-194.

[15] FaceIt Software Developer Kit, Identix Corporation, http://www.identix.com/

[16] D.M. Weinstein. "The Analytic 3-D Transform for the Least-Squared Fit of Three Pairs of Corresponding Points," School of Computing Technical Report, No. UUCS-98-005, University of Utah, Salt Lake City, UT, 1998.

[17] FaceVAC Software Developer Kit, Cognitec, http://www.cognitec-systems.de