3D Model-Based Face Recognition in Video

Unsang Park and Anil K. Jain

Department of Computer Science and Engineering
Michigan State University
3115 Engineering Building
East Lasing, MI 48824, USA
{parkunsa,jain}@cse.msu.edu

Abstract. Face recognition in video has gained wide attention due to its role in designing surveillance systems. One of the main advantages of video over still frames is that evidence accumulation over multiple frames can provide better face recognition performance. However, surveillance videos are generally of low resolution containing faces mostly in non-frontal poses. Consequently, face recognition in video poses serious challenges to state-of-the-art face recognition systems. Use of 3D face models has been suggested as a way to compensate for low resolution, poor contrast and non-frontal pose. We propose to overcome the pose problem by automatically (i) reconstructing a 3D face model from multiple non-frontal frames in a video, (ii) generating a frontal view from the derived 3D model, and (iii) using a commercial 2D face recognition engine to recognize the synthesized frontal view. A factorization-based structure from motion algorithm is used for 3D face reconstruction. The proposed scheme has been tested on CMU's Face In Action (FIA) video database with 221 subjects. Experimental results show a 40% improvement in matching performance as a result of using the 3D models.

Keywords: Face recognition, video surveillance, 3D face modeling, view synthesis, structure from motion, factorization, active appearance model.

1 Introduction

Automatic face recognition has now been studied for over three decades. While substantial performance improvements have been made in controlled scenarios (frontal pose and favorable lighting conditions), the recognition performance is still brittle with pose and lighting variations [12, 13]. Until recently, face recognition was mostly limited to one or more still shot images, but the current face recognition studies are attempting to combine still shots, video and 3D face models to achieve better performance. In particular, face recognition in video has gained substantial attention due to its applications in deploying surveillance systems. However, face images captured in surveillance systems are mostly off-frontal and have low resolution. Consequently, they do not match very well with the gallery that typically contains frontal face images.

S.-W. Lee and S.Z. Li (Eds.): ICB 2007, LNCS 4642, pp. 1085–1094, 2007.

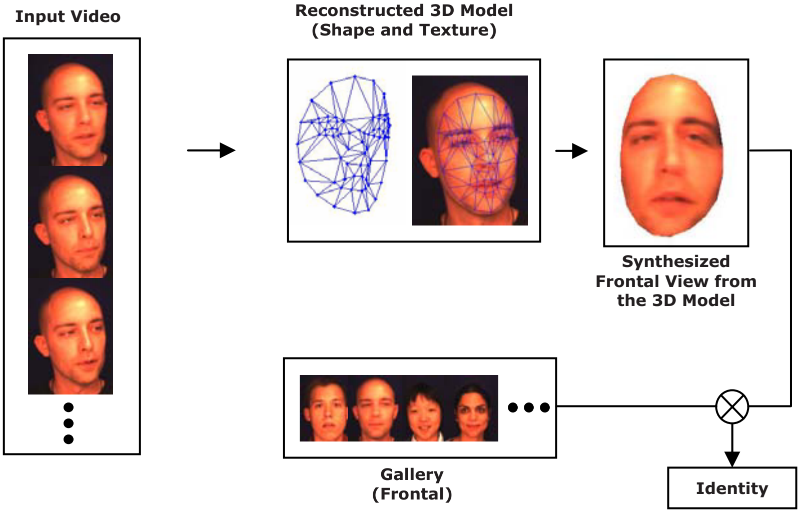


Fig. 1. Face recognition system with 3D model reconstruction and frontal view synthesis

There have been two main approaches to overcome the problem of pose and lighting variations: (i) view-based and (ii) view synthesis. View-based methods enroll multiple face images under various pose and lightings and match the probe image with the gallery image with most similar pose and lighting conditions [1, 2]. View-synthesis methods generate synthetic views from the input probe images with similar pose and lighting conditions as in gallery data to improve the matching performance. The desired view can be synthesized by learning the mapping function between pairs of training images [3] or by using 3D face models [4, 11]. The parameters of the 3D face model in the view synthesis process can also be used for face recognition [4]. Some of the other approaches for face recognition in video utilize appearance manifolds [14] or probabilistic models [15], but they require complicated training process and have been tested only on a small database.

The view-synthesis method is more appealing than the view-based method in two respects. First, it is not practical to collect face images at all possible pose and lighting conditions for the gallery data. Second, the state-of-the-art face recognition systems [5] perform the best in matching two near-frontal face images.

We propose a face recognition system that identifies the subject in a video which contains mostly non-frontal faces. We assume that only the frontal pose is enrolled in gallery data. This scenario is commonly observed in practical surveillance systems. The overall system is depicted in Fig. 1. One of the main contributions of the proposed work is to utilize 3D reconstruction techniques [6, 7, 8] for the purpose of handling the pose variation in the face recognition task. Unlike the morphable model based approach [4], comprehensive evaluations of 3D reconstruction from 2D images for the face recognition task have not been reported. Most of the effort in 3D face model reconstruction from 2D video has focused on accurate facial surface reconstruction, but the application of the resulting model for recognition has not been extensively explored. The contributions of our work are: (i) quantitative evaluation of the performance of factorization algorithm in structure from motion, (ii) view-synthesis using structure from motion and its use in face recognition, and (iii) evaluation on a public domain video database (CMU's Face In Action database [10]) using a commercial state-of-the-art face recognition engine (FaceVACS from Cognitec [5]).

2 3D Face Reconstruction

Obtaining a 3D face model from a sequence of 2D images is an active research problem. Morphable model (MM) [4], stereography [19], and Structure from Motion (SfM) [17, 6] are well known methods in 3D face model construction from 2D images or video. MM method has been shown to provide accurate reconstruction performance, but the processing time is overwhelming for use in real-time systems. Stereography also provides good performance and has been used in commercial applications [19], but it requires a pair of calibrated cameras, which limits its use in many surveillance applications. SfM gives reasonable performance, ability to process in real-time, and does not require a calibration process, making it suitable for surveillance applications. Since we are focusing on face recognition in surveillance video, we propose to use the SfM technique to reconstruct the 3D face models.

2.1 Tracking Facial Feature Points

We use 72 facial feature points that outline the eyes, eyebrows, nose, mouth and facial boundary. While the actual number of feature points is not critical, there needs to be sufficient number of points to capture the facial characteristics. Number of points used in face modeling using AAM vary in the range 60~100. The predefined facial feature points are automatically detected and tracked by the Active Appearance Model (AAM) [18], which is available as a SDK in public domain [9]. We train the AAM model on a training database with about ± 45° yaw, pitch and roll variations. As a result, the facial feature points from a face image within ±45° variations can be reliably located in the test images. The feature points detected in one frame are used as the initial locations for searching in the next frame, resulting in more stable point correspondences. An example of feature point detection is shown in Fig. 2.

2.2 3D Shape Reconstruction

The Factorization method [6] is a well known solution for the Structure from Motion problem. There are different factorization methods depending on the rigidity of the object [7, 8], to recover the detailed 3D shape. We regard the face as a rigid object and treat the small expression changes as noise in feature point detection, resulting in recovering only the most dominant shape from video data.

We use orthographic projection model that works reasonably well for face reconstruction when the distance between camera and object is a few meters. Under orthographic projection model, the relationship between 2D feature points and 3D shape is given by

$$W = M \cdot S, \tag{1}$$

$$W = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ & \vdots & & \\ u_{f1} & u_{f2} & \cdots & u_{fp} \\ v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ & \vdots & & \\ v_{f1} & v_{f2} & \cdots & v_{fp} \end{bmatrix},\ M = \begin{bmatrix} i_{1x} & i_{1y} & i_{1z} \\ i_{2x} & i_{2y} & i_{2z} \\ & \vdots & \\ i_{fx} & i_{fy} & i_{fz} \\ j_{1x} & j_{1y} & j_{1z} \\ j_{2x} & j_{2y} & j_{2z} \\ & \vdots & \\ j_{fx} & j_{fy} & j_{fz} \end{bmatrix},\ S = \begin{bmatrix} S_{x1} & S_{x2} & \cdots & S_{xp} \\ S_{y1} & S_{y2} & \cdots & S_{yp} \\ S_{z1} & S_{z2} & \cdots & S_{zp} \end{bmatrix}, \tag{2}$$

where, u_{fp} and v_{fp} in W represent the row and column pixel coordinates of p^th point in the f^th frame, each pair of $\boldsymbol{i_f^T} = [i_{fx}\ i_{fy}\ i_{fz}]$ and $\boldsymbol{j_f^T} = [j_{fx}\ j_{fy}\ j_{fz}]$ in M represents the rotation matrix with respect to the f^th frame, and S represents the 3D shape. The translation term is omitted in Eq. (1) because all 2D coordinates are centered at the origin. The rank of W in Eq. (2) is 3 in the ideal noise-free case.

The solution of Eq. (1) is obtained by a two-step process: (i) Find an initial estimate of M and S by singular value decomposition, and (ii) apply metric constraints on the initial estimates. By a singular value decomposition of W, we obtain

$$W = U{\cdot}D{\cdot}V^T \approx U'D'V'^T, \tag{3}$$

where U and V are unitary matrices of size 2F×2F and P×P, respectively and D is a matrix of size 2F×P for F frames and P tracked points. Given U, D and V, U' is the first three columns of U, D' is the first three columns and first three rows of D and V'^T is the first three rows of V^T, to impose the rank 3 constraint on W. Then, M' and S' that are the initial estimates of M and S are obtained as

$$\begin{aligned} M' &= U'{\cdot}D'^{1/2}, \\ S' &= D'^{1/2}{\cdot}V'^T. \end{aligned} \tag{4}$$

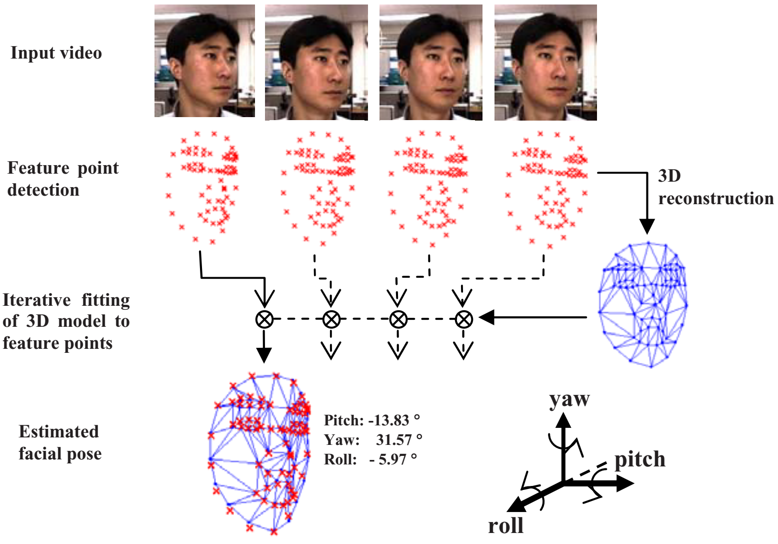


Fig. 2. Pose estimation scheme

To impose the metric constrains on *M'*, a 3×3 correction matrix *A* is defined as

$$([\boldsymbol{i}_f\ \boldsymbol{j}_f]^{\mathrm{T}} \cdot A) \cdot (A^T \cdot [\boldsymbol{i}_f\ \boldsymbol{j}_f]) = E, \tag{5}$$

where $\boldsymbol{i}_f$ is the f$^{\mathrm{th}}$ $\boldsymbol{i}$ vector in the upper half rows of *M*, $\boldsymbol{j}_f$ is the f$^{\mathrm{th}}$ $\boldsymbol{j}$ vector in the lower half rows of *M* and *E* is a 2×2 identity matrix. The constraints in Eq. (5) need to be imposed across all frames. There are one $\boldsymbol{i}_f$ and $\boldsymbol{j}_f$ vectors in each frame, which generate three constraints. Since $A \cdot A^T$ is a 3x3 symmetric matrix, there are 6 unknown variables. Therefore, at least two frames are required to solve Eq. (5). In practice, to obtain a robust solution, we need more than two frames and the solution is obtained by the least squared error method. The final solution is obtained as

$$\begin{aligned} M &= M' \cdot A, \\ S &= A^{-1} \cdot S', \end{aligned} \tag{6}$$

where *M* contains the rotation information between each frame and the 3D object and *S* contains the 3D shape information. We will provide the lower bound evaluation of the performance of Factorization method on synthetic and real data in Section 3.

2.3 3D Facial Pose Estimation

We estimate the facial pose in a video frame to select the best texture model for the 3D face construction. The video frame with facial pose close to frontal is a good candidate for texture mapping because it covers most of the face area. If a single profile image is used for texture mapping, the quality of the 3D model will be poor in the occluded region. When a frame in a near-frontal pose is not found, two frames are used for texture mapping. There are many facial pose estimation methods in 2D and 3D domains [20]. Because the head motion occurs in 3D domain, 3D information is necessary for accurate pose estimation. We estimate the facial pose in [yaw, pitch, roll] (YPR) values as shown in Fig. 2. Even though all the rotational relationships between the 3D shape and the 2D feature points in each frame are already obtained by the matrix *M* in factorization process, it reveals only the first two rows of the rotation matrix for each frame, which generates inaccurate solutions in obtaining YPR values especially in noisy data. Therefore, we use the gradient descent method to iteratively fit the reconstructed 3D shape to the 2D facial feature points.

2.4 Texture Mapping

We define the 3D face model as a set of triangles and generate a VRML object. Given the 72 points obtained from the reconstruction process, 124 triangles are generated. While the triangles can be obtained automatically by Delaunay triangulation process [16], we use a predefined set of triangles for efficiency sake because the number and configuration of the feature points are fixed. The corresponding set of triangles can be obtained from the video frames with a similar process. Then, the VRML object is generated by mapping the triangulated texture to the 3D shape. The best frame to be used in texture mapping is selected based on the pose estimation as described earlier. When all the available frames deviate

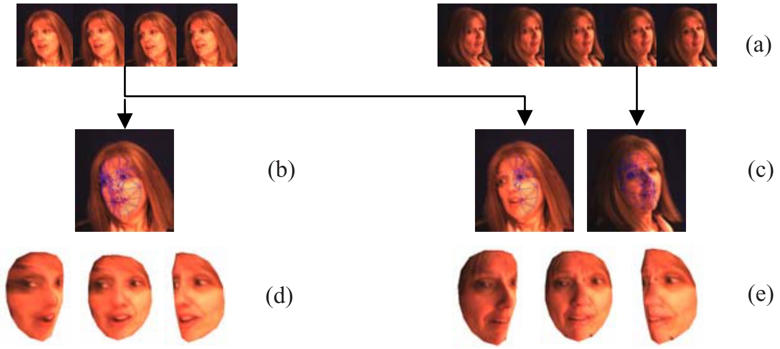


Fig. 3. Texture mapping. (a) a video sequence used for 3D reconstruction; (b) single frame with triangular meshes; (c) two frames with triangular meshes; (d) reconstructed 3D face model with one texture in (b); (e) reconstructed 3D face model with two texture mappings in (c). The two frontal poses in (d) and (e) are correctly identified in matching experiment.

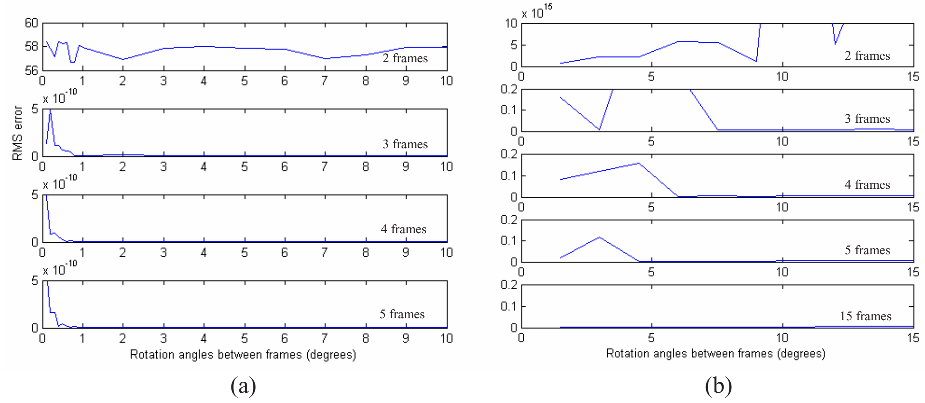
significantly from the frontal pose, two frames are used for texture mapping as described in Fig. 3. Even though both the synthetic frontal views in Figs. 3 (d) and (e) are correctly recognized, the one in (e) looks more realistic. When more than one texture is used for texture mapping, a sharp boundary is often observed across the boundary where two different textures are combined because of the difference in illumination. However, the synthetic frontal views are correctly recognized in most cases regardless of this artifact.

3 Experimental Results

We performed a number of experiments to i) evaluate the lower performance bound of the Factorization algorithm in terms of the rotation angle and number of frames in synthetic and real data, ii) reconstruct 3D face models on a large public domain video database, and iii) perform face recognition using the reconstructed 3D face models.

3.1 3D Face Reconstruction with Synthetic Data

We first evaluate the performance of the Factorization algorithm using synthetic data. A set of 72 facial feature points are obtained from a ground truth 3D face model. A sequence of 2D coordinates of the facial feature points are directly obtained from the ground truth. We take the angular values for the rotation in steps of 0.1° in the range (0.1°, 1°) and in steps of 1° in the range (1°, 10°). The number of frames range from 2 to 5. The RMS error between the ground truth and the reconstructed shape is shown in Fig. 4 (a). While the number of frames required for the reconstruction in the noiseless case is two (Sec. 2.2), in practice more frames are needed to keep the error small. As long as the number of frames is more than two, the errors are negligible.

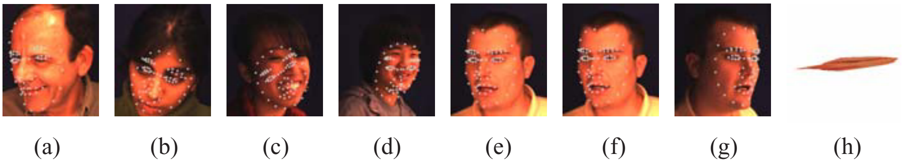


(a) (b)

Fig. 4. (a) RMS error between reconstructed shape and ground truth. (b) RMS error between reconstructed and ideal rotation matrix, M_s.

3.2 3D Face Reconstruction with Real Data

For real data, noise is present in both the facial feature point detection and the correspondences between detected points across frames. This noise is not random and its affect is more pronounced at points of self-occlusion and on the facial boundary. Since AAM does use feature points on the facial boundary, the point correspondences are not very accurate in the presence of self-occlusion. Reconstruction experiments are performed on real data with face rotation from -45° to +45° across 61 frames. We estimate the rotation between successive frames as 1.5° (61 frames varying from -45° to +45°) and obtain the reconstruction error with rotation in steps of 1.5° in the range (1.5°, 15°). The number of frames used is from 2 to 61. A direct comparison between the ground truth and the reconstructed shape is not possible in case of real data because the ground truth is not known. Therefore, we measure the orthogonality of M to estimate the reconstruction accuracy. Let M be a 2F×3 matrix as shown in Eq. (3) and $M(a{:}b,c{:}d)$ represent the sub matrix of M from rows a to b and columns c to d. Then, $M_s = M \times M'$ is a 2Fx2F matrix where all elements in $M_s(1{:}F,\ 1{:}F)$ and $M_s(F+1{:}2F,\ F+1{:}2F)$ are equal to 1 and all elements in $M_s(1{:}F,\ F+1{:}2F)$ and $M_s(F+1{:}2F,\ 1{:}F)$ are equal to 0 if M is truly an orthogonal matrix. We measure the RMS difference between the ideal M_s and the calculated M_s as the reconstruction error. The reconstruction error for real data is shown in Fig. 4 (b).



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5. Examples where 3D face reconstruction was not successful. (a), (b), (c) and (d) Failure of feature point detection using AAM; (e), (f) and (g) Deficiency of motion cue, resulting in (h) a failure of SfM to reconstruct the 3D face model.

Based on experiments with real data, it is observed that the number of frames needed for reconstruction is more for real data than the synthetic data, but the error decreases quickly as the number of frames increases. The slight increase in error with larger pose differences is due to error in point correspondences from self-occlusion.

3.3 Face Recognition with Pose Correction

We have used a subset of CMU's Face In Action (FIA) video database [10] that includes 221 subjects for our matching experiments. To demonstrate the advantage of using reconstructed 3D face models for recognition, we are primarily interested in video sequences that contain mostly non-frontal views for each subject. Since the reconstruction with SfM performs better when there are large motions, both left and right non-frontal views are collected for each subject in FIA database, if available, resulting, on average, about 10 frames per subject. When there is a sufficient inter-frame motion and the feature point detection performs well, it is possible to obtain the 3D face model from only 3 different frames, which in consistent with the results shown in Fig. 4. The number of frames that is required for the reconstruction can be determined based on the orthogonality of *M*.

We successfully reconstructed 3D face models for 197 subjects out of the 221 subjects in the database. The reconstruction process failed for 24 subjects either due to poor facial feature point detection in the AAM process or the deficiency of motion cue, which caused a degenerate solution in the factorization algorithm. Example images where AAM or SfM failed are shown in Fig. 5. The failure occurs due to large pose or shape variations that were not represented in the samples used to train the AAM model. All the reconstructed 3D face models are corrected in their pose to make all yaw, pitch, and roll values equal to zero. The frontal face image can be obtained by projecting the 3D model in the 2D plane. Once the frontal view is synthesized, FaceVACS® face recognition engine from Cognitec [5] is used to generate the matching score. This engine is one of the best commercial 2D face recognition systems. The face recognition results for frontal face video, non-frontal face video

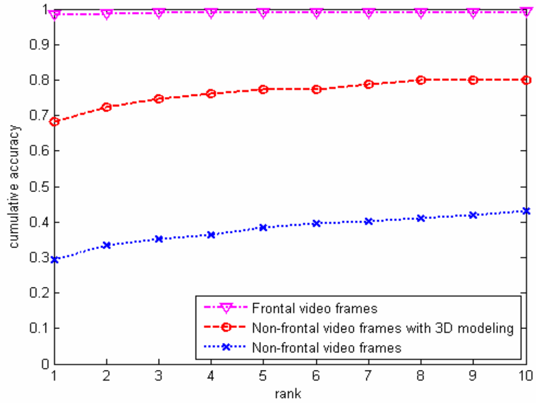


Fig. 6. Face recognition performance with 3D face modeling

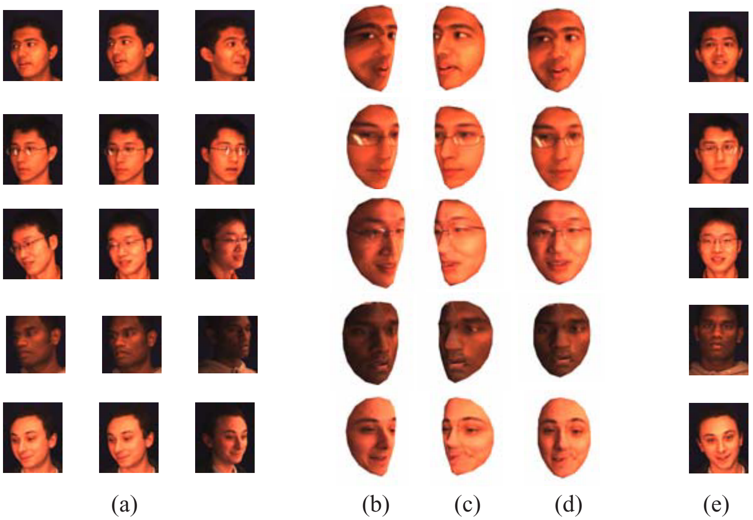


Fig. 7. 3D model-based face recognition results on six subjects (Subject IDs in the FIA database are 47, 56, 85, 133, and 208). (a) Input frames; (b), (c) and (d) reconstructed 3D face models at right, left, and frontal views, respectively; (e) frontal images enrolled in the gallery database. All the frames in (a) are not correctly identified, while the synthetic frontal views in (d) obtained from the reconstructed 3D models are correctly identified for the first four subjects, but not for the last subject (#208). The reconstructed 3D model of the last subject appears very different from the gallery image, resulting in the recognition failure.

and non-frontal face video with 3D face modeling are shown in Fig. 6 based on 197 subjects for which the 3D face reconstruction was successful. The CMC curves show that the FaceVACS engine does extremely well for frontal pose video but its performance drops drastically for non-frontal pose video. By using the proposed 3D face modeling, the rank-1 performance in non-frontal scenario improves by about 40%. Example 3D face models and the synthesized frontal views from six different subjects are shown in Fig. 7.

4 Conclusions

We have shown that the proposed 3D model based face recognition from video provides a significantly better performance, especially when many non-frontal views are observed. The proposed system synthesizes the frontal face image from the 3D reconstructed models and matches it against the frontal face enrolled in the gallery. We have automatically generated 3D face models for 197 subjects in the Face In Action database (session 1, indoor, camera 5) using the SfM factorization method. The experimental results show substantial improvement in the rank-1 matching performance (from 30% to 70%) for video with non-frontal pose. The entire face recognition process (feature point tracking, 3D model construction, matching) takes ~10 s per subject on a Pentium IV PC for 320x240 frames. We are working to improve the matching speed and the texture blending technique to remove the sharp boundary that is often observed when mapping multiple textures.

References

[1] Pentland, A., Moghaddam, B., Starner, T.: View-based and Modular Eigenspace for Face Recognition. In: Proc. CVPR, pp. 84–91 (1994)

[2] Chai, X., Shan, S., Chen, X., Gao, W.: Local Linear Regression (LLR) for Pose Invariant Face Recognition. In: Proc. AFGR, pp. 631–636 (2006)

[3] Beymer, D., Poggio, T.: Face Recognition from One Example View. In: Proc. ICCV, pp. 500–507 (1995)

[4] Blanz, V., Vetter, T.: Face Recognition based on Fitting a 3D Morphable Model. IEEE Trans. PAMI 25, 1063–1074 (2003)

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

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

[7] Xiao, J., Chai, J., Kanade, T.: A Closed-Form Solution to Non-Rigid Shape and Motion Recovery. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 668–675. Springer, Heidelberg (2004)

[8] Brand, M.: A Direct Method for 3D Factorization of Nonrigid Motion Observation in 2D. In: Proc. CVPR, vol. 2, pp. 122–128 (2005)

[9] Stegmann, M.B.: The AAM-API: An Open Source Active Appearance Model Implementation. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2879, pp. 951–952. Springer, Heidelberg (2003)

[10] Goh, R., Liu, L., Liu, X., Chen, T.: The CMU Face In Action (FIA) Database. In: Zhao, W., Gong, S., Tang, X. (eds.) AMFG 2005. LNCS, vol. 3723, pp. 255–263. Springer, Heidelberg (2005)

[11] Zhao, W., Chellappa, R.: SFS Based View Synthesis for Robust Face Recognition. In: Proc. FGR, pp. 285–292 (2000)

[12] Phillips, P.J., Grother, P., Micheals, R.J., Blackburn, D.M., Tabassi, E., Bone, J.M.: FRVT: 2002: Evaluation Report, Tech. Report NISTIR 6965, NIST (2003)

[13] Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Worek, W.: Preliminary Face Recognition Grand Challenge Results. In: Proc. AFGR, pp. 15–24 (2006)

[14] Lee, K., Ho, J., Yang, M., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. CVPR I, 313–320 (2003)

[15] Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding 91, 214–245 (2003)

[16] Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The Quickhull Algorithm for Convex Hulls. ACM Trans. Mathematical Software 22(4), 469–483 (1996)

[17] Ullman, S.: The Interpretation of Visual Motion. MIT Press, Cambridge, MA (1979)

[18] Matthews, I., Baker, S.: Active Appearance Models Revisited. International Journal of Computer Vision 60(2), 135–164 (2004)

[19] Maurer, T., Guigonis, D., Maslov, I., Pesenti, B., Tsaregorodtsev, A., West, D., Medioni, G.: Performance of Geometrix ActiveIDTM 3D Face Recognition Engine on the FRGC Data. In: Proc. CVPR, pp. 154–160 (2005)

[20] Tu, J., Huang, T., Tao, H.: Accurate Head Pose Tracking in Low Resolution Video. In: Proc. FGR, pp. 573–578 (2006)