3D Model-Assisted Face Recognition in Video

Unsang Park, Hong Chen, and Anil K. Jain
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{parkunsa, chenhon2, jain}@cse.msu.edu

Abstract

Face recognition in video has gained wide attention as a covert method for surveillance to enhance security in a variety of application domains (e.g., airports). A video contains temporal information as well as multiple instances of a face, so it is expected to lead to better face recognition performance compared to still face images. However, faces appearing in a video have substantial variations in pose and lighting. These pose and lighting variations can be effectively modeled using 3D face models. Combining the advantages of 2D video and 3D face models, we propose a face recognition system that identifies faces in a video. The system utilizes the rich information in a video and overcomes the pose and lighting variations using 3D face model. The description of the proposed method and preliminary results are provided.

1. Introduction

Face recognition has been well studied with (2D) still images for over a decade. In still image based face recognition systems, a snapshot of a user is acquired and compared with gallery faces to establish a person's identity. In this procedure, the user is expected to be cooperative to provide a frontal face image under uniform lighting conditions to enable the capture of a high quality face image. However, it is now well known that even small variations in pose and lighting can drastically degrade the performance of the single-shot 2D image based face recognition systems. Pose and lighting invariant face recognition is a challenging research area and various approaches have been proposed.

The two most well known approaches for achieving pose and lighting invariant face recognition are based on utilizing a 2D (containing multiple images and temporal information) video [14], [15] or 3D face model (containing surface geometry information) [7], [8], [11]. A video provides multiple face images (of the same person) [1], [2] as well as temporal information (e.g., movements of facial features) that can be used to improve face recognition performance. Given the trajectories of facial feature movement, face recognition is performed based on the similarities of the trajectories [3]. The trajectories can also be captured as nonlinear manifolds and the distance between clusters of faces in the feature space establishes the identity associated with the face [4]. The face recognition scenarios that use one or more 2D images are summarized in Table 1 [1].

Table 1. Face recognition scenarios for 2D images.

probe \ gallery	single still image	many still images	video
single still image	one-to-one	many-to-one	video-to-still
many still images	one-to-many	many-to-many	video-to-many
video	one-to-video	many-to-video	video-to-video

3D 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 [5]. Table 2 extends Table 1 across 2D face models.

Table 2. Face recognition settings across 2D and 3D inputs.

probe \ gallery	2D image	3D model
2D image	2D to 2D	2D to 3D
3D model	3D to 2D	3D to 3D

There have been many studies on 3D face recognition using reconstructed 3D models from a set of 2D images [6], [7]. The reconstructed 3D model is used to obtain the 2D projection images that are matched with probe images [5]. Alternatively, the reconstructed 3D model can be used to generate a frontal view of the probe image with arbitrary pose and lighting and then the recognition is performed with the synthesized frontal faces. Figure 1 shows a 3D model and its corresponding 2D projection images under pose and lighting conditions.

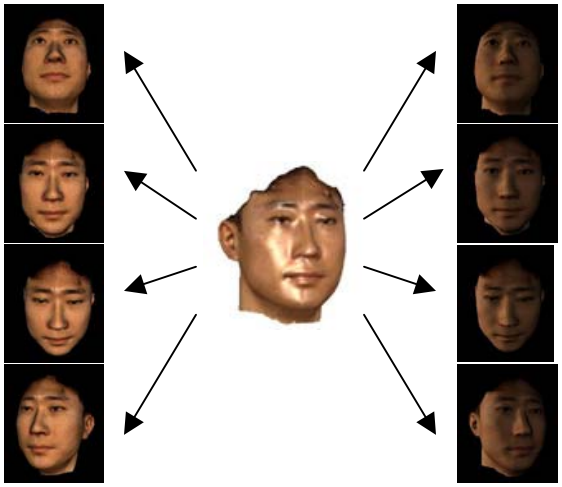


Figure 1. A 3D model and its 2D projections.

A combination of 2D and 3D face recognition systems is also regarded as a promising method [8],[11]. Typical (2D + 3D) methods match intensity data to intensity data and range data to range data [8], which means the 3D model needs to be acquired both at enrollment stage and identification stage. A 2D and 3D mixed matching is also described in [8], where the 2D projection images of the 3D model are used to construct the subspace of LDA for 2D face recognition.

We propose a face recognition system that automatically determains the identity of a person (say, on a watch list) in the video by utilizing her 3D face model. The system uses the images in the video as probe images and the identity is determined by comparing the probe images with the 2D projection images of the gallery 3D model under varying pose and lighting conditions. The proposed system has several advantages over existing approaches: (i) it utilizes the video that contains multiple face images of a subset and temporal information, (ii) 3D model can effectively handle the pose and lighting variations, (iii) by using the true 3D model, potential errors in reconstructing 3D model from 2D images are avoided, and (iv) 3D face scans are not required in the recognition stage because the probe (2D) image is matched against 2D projections of the enrolled 3D model.

The paper is composed as follows. Section 2 describes the probe and gallery data for our face recognition system. Section 3 introduces an automatic method of estimating the pose and lighting conditions of a face extracted from the video. Section 4 introduces a description of our recognition scheme. Section 5 describes the experimental results and section 6 concludes this paper.

2. Probe and Gallery Data

2.1. Probe Data

Ten video files are recorded for ten subjects under four different lighting conditions at various poses with yaw and pitch motion. Even though the eventual system is being targeted to process raw video with arbitrary pose and lighting, the current system is tested on a subset of pose and lighting variations captured from the video for the design and evaluation purposes. 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 2.

Figure 2. Pose and lighting variations in a video.

2.2. Gallery Data

One hundred 3D face models (of 100 different subjects) available in our laboratory were used to generate the gallery images. A Minolta VIVID 910 laser scanner was used to acquire the texture and range data of the faces of 100 different subjects from 5 different angles: frontal, right 30, right 60, left 30, and left 60 degrees. These five 2.5D scans were then stitched together to construct a full 3D model of a face. Some screenshots of 3D face models of 5 different subjects are shown in Figure 3.

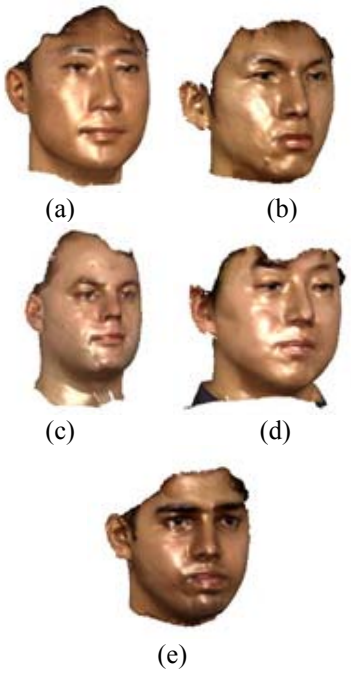


Figure 3. 3D models for 5 subjects.

The 3D face models are converted to Virtual Reality Modeling Language (VRML) [9] objects that provide rich control and rendering options for the model. The VRML objects are carefully controlled with varying pose and lighting conditions and 2D projections of the 3D face models are generated to form the gallery images. Sample gallery images generated from the 3D model of one subject (shown in Figure 1) are shown in Figure 4.

3. Pose and Lighting Estimation

Pose and lighting estimation is often required in objection recognition from video [3] either to focus on a certain pose or select a certain sequence of poses. We use a sequence of images from video for the recognition purpose using temporal cue. Preliminary results on automatic pose and lighting estimation are provided here.

Figure 4. 2D projections of the 3D face model of a subject with the same pose and lighting variations as in Figure 1.

We quantize the pose and lighting to be a discrete set and simplify it as a classification problem with a predefined number of classes. We employed normal, dark, and 45 and 90 degrees from front to right as the set of lighting classes and frontal, right, left, up, and down as the set of pose classes. A support vector machine is used to classify the individual frames in the video into one of these lighting and pose classes. Traditional SVMs solve two-class classification problems, and they do not provide posterior probabilities. To solve multi-class classification problems, a strategy of one-versus-one or a strategy of one-versus-all is used [12]. One-versus-one classifiers are typically less complex than one-versus-all classifiers. Therefore, the former can be trained with smaller data sets. On the other hand, if there are M classes, $M(M-1)/2$ SVMs are needed for the one-versus-one strategy, as compared to M SVMs for the strategy of one-versus-all. Experiments on a number of standard classification tasks have shown that one-versus-one classifiers are marginally more accurate than one-versus-all classifiers [13]. Therefore, our system uses the strategy of one-versus-one.

The algorithm for pose and lighting estimation is summarized as follows:

1. Face detection and vectorization. The face is extracted from individual frames, and it is

normalized to a fixed size (20×20). The normalized image region is then vectorized for SVM training and testing.

2. Training SVMs. Half of the subjects are selected randomly from the database, and their face vectors are used for training SVMs.
3. Testing. The face images of the remaining subjects are used for testing purpose.

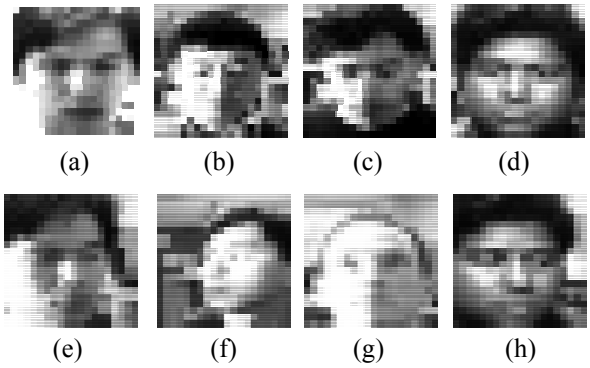


Figure 5. Correct and incorrect lighting estimations. (a), (b), (c) and (d) are correct estimations of lighting as normal, 45°, 90° and dark. (e), (f), (g) and (h) show misclassification of 45°, 45°, 90° and dark as 90°, dark, 45° and 45°, respectively.

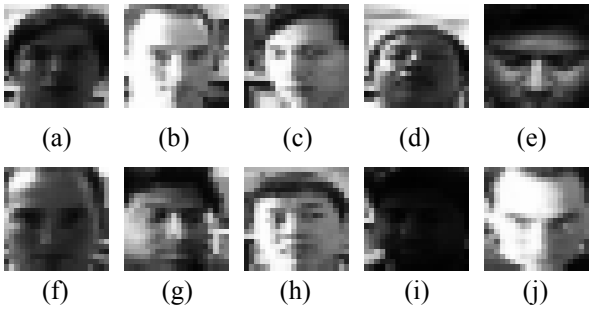


Figure 6. Correct and incorrect pose estimations. (a), (b), (c), (d) and (e) are correct estimations of pose as frontal, left, right, up and down. (e), (f), (g) and (h) show misclassifications of frontal, left, right, right and down as down, down, front, down and frontal, respectively.

In the experiments, 5 subjects were randomly selected from the database for training, and the remaining 5 subjects were used for testing. This procedure is repeated 5 times for the purpose of cross-validation. Since we have approximately 100 images per subject, the average number of testing images is 498. Figure 5 shows some examples of correct and incorrect lighting estimations. Figure 6 shows some examples of correct and incorrect pose classifications. The 4-class recognition accuracy for lighting estimation is 88.7%, and the recognition accuracy for 5-class pose estimation is 76.5%. The average confusion matrices are shown in tables 3 and 4.

Table 3. Confusion matrix of lighting estimation.

Prediction / label	Dark	45°	90°	Normal
Dark	125	0	0	0
45°	0	116.2	21.2	1.6
90°	0	21.8	87.8	0
Normal	0	7	5	113

Table 4. Confusion matrix of pose estimation.

Prediction / label	Frontal	Left	Right	Up	Down
Frontal	64.4	5.8	24.2	6.2	1.4
Left	2.4	74.2	7	13.4	2
Right	1.6	0.8	95.2	1.6	0.4
Up	3.0	1.8	2.6	92.6	0
Down	7.4	3.2	28.4	3.8	55.2

4. Recognition

For the recognition, face and facial features (e.g., eye) need to be detected first. Some examples of correct and incorrect face and eye detection from the probe image are provided in Figure 7 using FaceIt® SDK from Identix [10].

Let $G = \{G_i, i=1,2,\ldots,N_G\}$ be the set of 3D models enrolled as the gallery data. Given a probe image p, the identity is decided by

$$ID = \arg \max_{i=1,2,\ldots,N_g} s(p, G_i), \quad (1)$$

where $s(\cdot)$ represents the matching score measured by FaceIt® SDK. Having a set of 2D projection images $g_i = \{g_{ij}, j=1,2,\ldots,N_{gi}\}$ for each model G_i, the identity can be equivalently decided by

$$ID = \arg \max_{i=1,2,\ldots,N_g} s(p, g_i). \quad (2)$$

To utilize the advantage of input video, recognition using multiple images and temporal cue is explored. Majority voting and score sum are used to fuse the recognition result from multiple frames. Temporal cue is used in the sense of matching the sequence of matching scores from a sequence of frames.

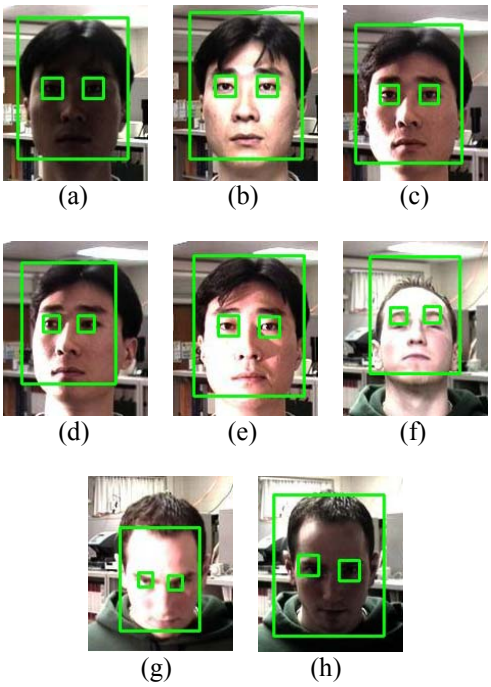


Figure 7. Sample images of correct and incorrect face and eye detections. (a), (b), (c), (d) and (e) show correct examples. (f), (g) and (h) show incorrect examples.

In majority voting, the maximum matching scores are decided for a set of probe images q = $\{p_j, j=1,2,\ldots,N_w\}$ as

$$ID_j = \arg \max_{i=1,2,\ldots,N_g} s(p_j, g_i), \forall j, \qquad (3)$$

then, the identity is decided from the majority *ID* in the set of probe images as

$$ID_{mv} = majority\{ID_j, j = 1,2,\ldots,w\}. \qquad (4)$$

In score sum, the matching scores are summed up for a set of probe images q = $\{p_j, j=1,2,\ldots,N_w\}$, then the identity is decided from the maximum of the sum of matching scores as

$$ID_{ss} = \arg \max_{i=1,2,\ldots,N_g} \sum_j s(p_j, g_i). \qquad (5)$$

To use temporal cues for the recognition, a LDA based classifier is used. After the face pose in a video is estimated, frames of different poses under specific lighting condition and specific order are extracted to form a probe sequence. The 2D projections of 3D model in the gallery under the same lighting condition are extracted in the same order to form a gallery sequence. The order of the poses we used in this experiment are right, up, left, down and frontal. A 5×5 matrix is generated from the set of matching scores of a probe sequence and a gallery sequence. Maximum values in all the rows in the matrix are extracted to establish a 5-element feature vector X, and a weighted linear combination of the 5 values in the vector is used for ranking the subjects in the database. The weight vector W is obtained by Fisher discriminant analysis and W^TX is used for ranking. We applied the leave-one-out strategy in our experiment. One of the 10 subjects in the database is used as the test subject, and the remaining 9 subjects are used as training subjects. The diagram of the overall system is shown in Figure 8.

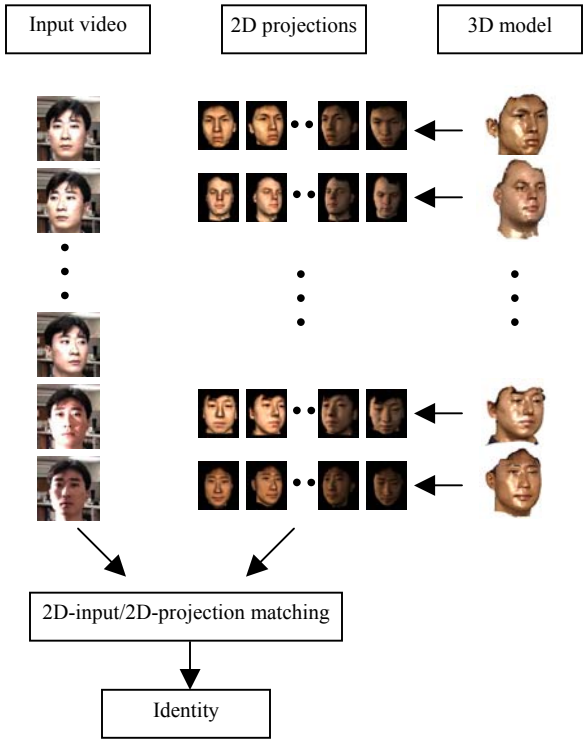


Figure 8. System diagram of the face recognition system using video input and 3D model gallery.

5. Experimental Results

Figure 9 shows the rank-N accuracies of face identification test with and without pose and lighting variations. The baseline 2D face matcher (from Identix) performs very well when both probe and gallery faces are frontal under normal lighting condition. However, the performance drops severely when the probe images have pose and lighting variations, and this is the main problem that we address in this work.

Figure 10 shows the rank-N accuracy with various sets of enrolled gallery images per person. The gallery database consists of one frontal 2D projection, one projection with the same pose and/or lighting with the

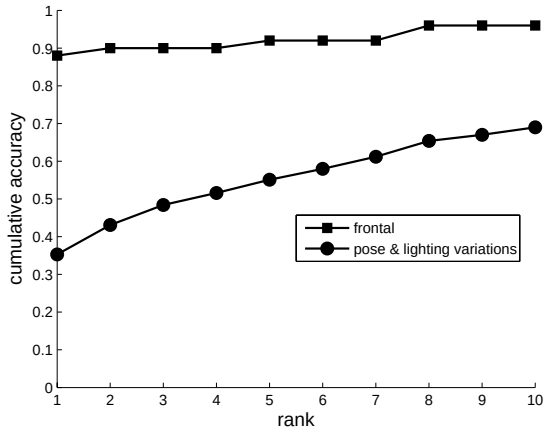


Figure 9. Rank-N accuracy with and without pose and lighting variations in the probe data.

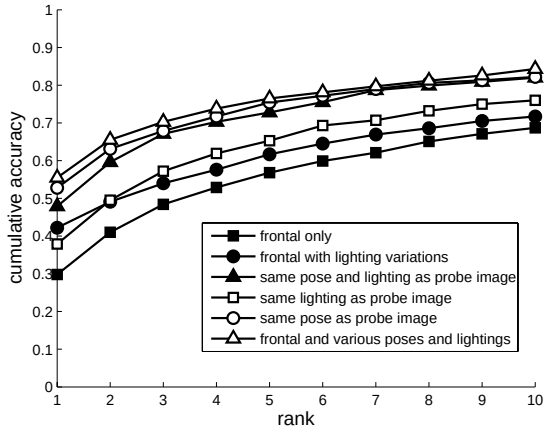


Figure 10. Rank-N accuracy with various gallery data.

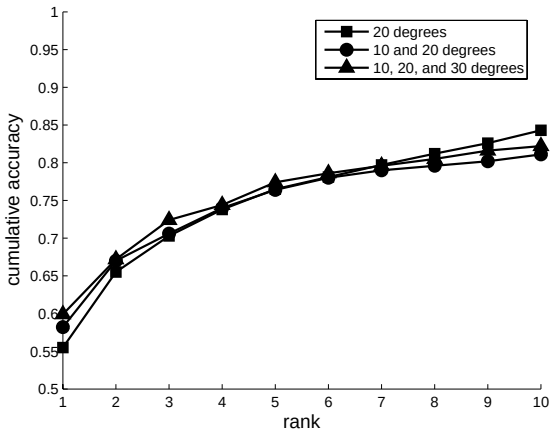


Figure 11. Rank-N accuracy with pose variations in the gallery.

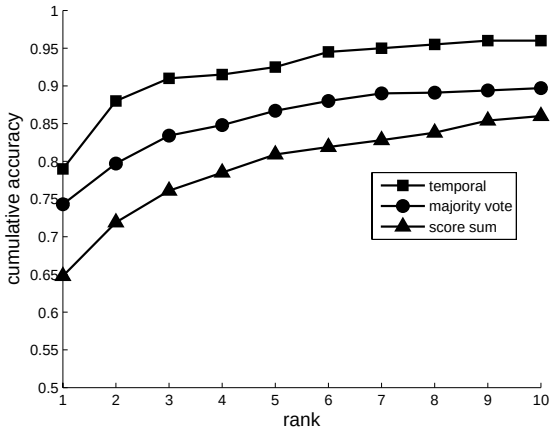


Figure 12. Rank-N accuracy using fusion rules with five frames and temporal cue in the same number of frames.

probe image, 20 2D projections with various pose and lightings per subject. The pose and lighting variations in the gallery data improves the recognition accuracy. Figure 11 also shows the effect of gallery data on the face recognition performance with respect to an increasing number of poses. Once the gallery data has certain level of pose and lighting variations, additional pose variations in the gallery data gave a marginal improvement in the recognition accuracy. Figure 12 shows the performance of using multiple images with fusion rule and temporal cue. Given 5 sampled frames of different poses, using temporal cue provided highest accuracy. Figure 13 shows the effect of number of

frames on the accuracy in the majority voting scheme. More number of frames provides better performance monotonically.

Some sample matching results are shown in Figure 14. A majority of the errors occur in non-frontal pose. An interesting observation is that the incorrect matches mostly happen in the same pose. This shows that the intra-class variability (appearance differences in different poses for the same subject) exceeds the inter-class variability.

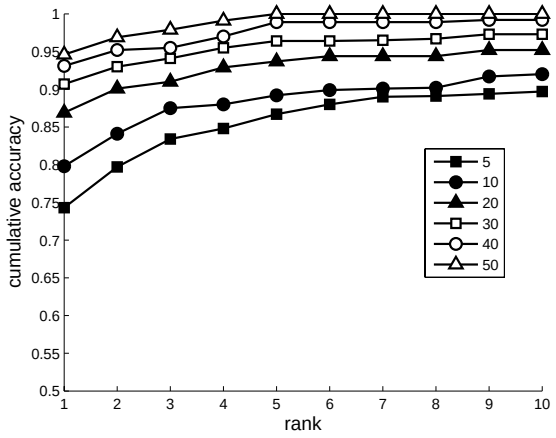
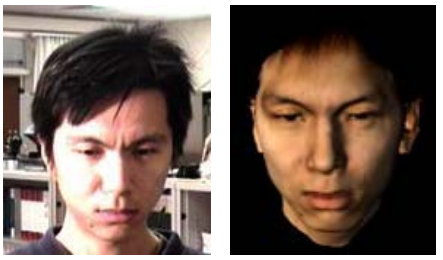
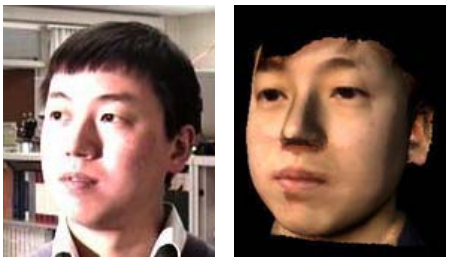


Figure 13. Rank-N accuracy using fusion rules with different numbers of frames.

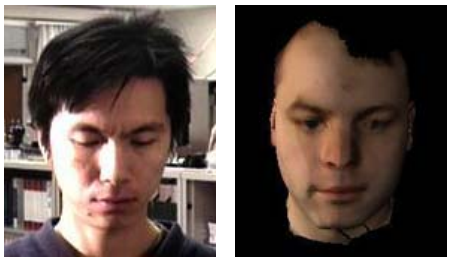


(a)

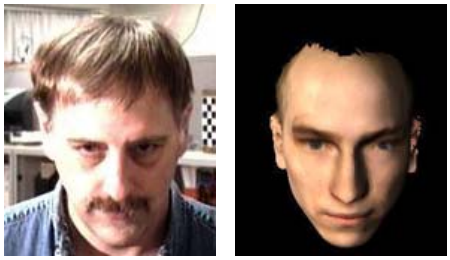


(b)

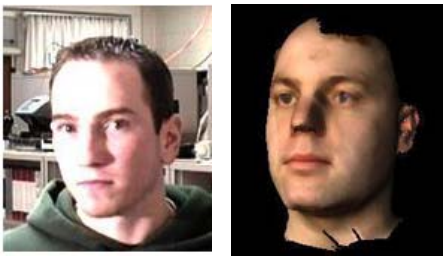
(c)



(d)



(e)



(f)

Figure 14. Sample images of correct and incorrect matches. Each pair shows the input image and the rank-1 match of 2D projection. (a), (b) and (c) are correct matches. (d), (e) and (f) are incorrect matches.

6. Conclusions and Future Work

We have proposed a face recognition system that uses video as input and 3D model as gallery. Our proposed matching scheme based on multiple 2D projection images from 3D model is shown to be better than single 2D image-based recognition. By taking advantage of the video and 3D model, our proposed

method overcomes various problems in current face recognition systems. However, there are certain limitations in the proposed system. The system has been tested only on a small variations of pose and lightings and does not perform well in uncontrolled environments. This may be an inherent limitation of the 2D-based face recognition algorithm. On the other hand, the problem may be due to other factors such as the distortion of images with varying distance of the subject to camera or the improper selection of frames in the video.

The current face identification scenario takes about 2~3 seconds with 2000 images in the gallery. The execution time is acceptable for a general face identification system, but not suitable for a real-time surveillance system. The process can be made faster by an advanced indexing method that quickly trims down the gallery to a small set using some distinctive features of the probe image such as skin color or high-level facial features.

Future work will include fully automating the recognition system from the video input to identification, improving the system to work under uncontrolled general situation, and developing an advanced indexing scheme for fast process. More experiments on direct 2D/3D matching in 2D/3D domain will also be performed in the future.

7. References

[1] S. K. Zhou, "Face recognition using more than one still image: What is more?," Proc. 5th Chinese Conference on Biometric Recognition, (SINOBIOMETRICS), Guangzhou, pp. 225-232, 2004.

[2] F. Fraser, "Exploring the use of face recognition technology for border control application – Australia's experience," Biometric Consortium Conference, 2003.

[3] Y. Li, S. Gong, and H. Liddell, "Video-based online face recognition using identity surfaces," Proc. IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Vancouver, Canada, pp. 40-46, 2001.

[4] O. Arandjelović and R. Cipolla, "An illumination invariant face recognition system for access control using video," Proc. British Machine Vision Conference, 2004.

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

[6] 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, pp. 277-284, 2000.

[7] Volker Blanz and Thomas Vetter**, "**Face recognition based on fitting a 3D morphable model**,"** IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 9, pp. 1063-1074, 2003.

[8] Xiaoguang Lu and Anil K. Jain, "Integrating Range and Texture Information for 3D Face Recognition," Proc. IEEE WACV, Breckenridge, Colorado, 2005.

[9] Jed Hartman and Josie Wernecke, "The VRML 2.0 Handbook: Building Moving Worlds on the Web, "Addison-Wesley, ISBN 0-201-47944-3, 1996.

[10] http://www.identix.com/

[11] K. Chang, K. Bowyer, and P. Flynn, "Face recognition using 2D and 3D facial data," ACM Workshop on Multimodal User Authentication, pp. 25-32, 2003.

[12] John C. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in Advances in large Margin Classifiers, Alexander J. Smola, Peter Bartlett, Bernhard Schlkopf, and Dale Schuurmans, Eds., MIT Press, Cambridge, MA, 1999.

[13] E. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," J. Machine Learning Research, vol. 1, pp. 113–141, 2000.

[14] R. S. Feris, T. E. Campos and R. M. Cesar Jr. A Project for Face Recognition from Video Sequences Using GWN and Eigenfeature Selection. In Proceedings of WAICV'2000 Workshop on Artificial Intelligence and Computer Vision, pp. 141-145, Atibaia, Brazil, November 2000

[15] G. Gordon, M. Lewis, ``Face Recognition Using Video Clips and Mug Shots", Proceedings of the Office of National Drug Control Policy (ONDCP) International Technical Symposium (Nashua, NH), October 1995