Face Recognition by Multi-Frame Fusion of Rotating Heads in Videos

Shaun J. Canavan, Michael P. Kozak, Yong Zhang, John R. Sullins,

Matthew A. Shreve and Dmitry B. Goldgof

Abstract

This paper presents a face recognition study that implicitly utilizes the 3D information in 2D video sequences through multi-sample fusion. The approach is based on the hypothesis that continuous and coherent intensity variations in video frames caused by a rotating head can provide information similar to that of explicit shapes or range images. The fusion was done on the image level to prevent information loss. Experiments were carried out using a data set of over 100 subjects and promising results have been obtained: (1) under regular indoor lighting conditions, rank one recognition rate increased from 91% using a single frame to 100% using 7-frame fusion; (2) under strong shadow conditions, rank one recognition rate increased from 63% using a single frame to 85% using 7-frame fusion.

I. Introduction

3D face recognition has received much attention recently in the biometrics research community, because 3D faces are considered to be less affected by the illumination and pose variations that often plague the 2D image-based approaches. Almost all studies have shown improvements in recognition accuracy when 2D and 3D faces were combined. However, as pointed out by Bowyer *et al* [1] and Kakadiaris et al [2], the use of 3D shapes, especially range images, has a few limitations: (1) current sensors have limited operation ranges (< 2m); (2) 3D data require much more storage space and long processing time; (3) acquisition is often not fully automated and may need user intervention. It is unlikely that those technical issues will be completely resolved in the near future. Therefore, there is a strong interest to explore other methods or sources that can provide 3D information that is equivalent or complementary to that of range images.

In this paper, we propose a method that utilizes a video sequence in which a subject gradually rotated his/her head from the frontal view to the profile view. Our hypothesis is that the 3D geometry of a rotating face should be embedded in the continuous intensity changes of an image stream, and therefore can be harnessed by the recognition algorithm without the need of an explicit 3D face model. Multiple video frames that capture the face at different pose angles can be combined to provide a more reliable and comprehensive 3D representation of the face than any single view image. The proposed method has several advantages:

1) If the video sequences acquired by a regular or high definition camcorder can provide quality 3D data for face recognition, some of the constraints posed by 3D sensors can be relieved. For example, an optical camcorder has a much wider operation range and can record videos in real time. Therefore, this method has the potential to be deployed in realistic settings such as access control, security checking and video surveillance.
2) Since the 3D information of a face is implicitly inferred by multi-frame fusion, the high computational cost of explicit 3D modeling (whether using a surface mesh or a solid mesh) can be avoided.
3) Not all frames in a rotating head video will be used (the number of frames can reach a few hundreds in a rotation sequence). Fusion of a few selected frames often suffices the need of face recognition. This gives us the flexibility to perform fusion either on the image or score level.
4) More importantly, a video sequence of a face with different poses might help alleviate the adverse effect of lighting changes on recognition accuracy. For instance, a light source can cast shadows on a face, but at the same time, it also reveals the 3D curvatures of the face by creating sharp intensity contrasts (such as silhouette).

II. Related Works

A literature survey with in-depth discussions of the current developments in 3D methods can be found in [1]. Zhao *et al* [3] presented a more extensive survey of the existing methods in face recognition. In this section, we give a brief review of the techniques that are most relevant to our approach.

One motivation of using videos of rotating heads to facilitate face recognition is that, a few recent studies have demonstrated that the multi-sample approach can achieve a performance comparable to that of the multi-modal approach. Using a data set of varying facial expressions and lighting conditions, Bowyer *et al* [4] reported an improvement in rank one recognition rate from 96.1% with two frames per subject to 100% with four frames per subject. In another study, Thomas *et al* [5] observed that the recognition rate generally increases as the number of frames per subject increases, regardless of the type of camera being used. They also found that the optimal number of frames per subject is between 12 and 18, given the particular data sets used. However, as noted

This work was supported in part by Ohio Board of Regents Research Incentive Grant No. 34241 and Youngstown State University Research Council Grants No. 07-#16 and No. 08-#8.

S. J. Canavan, M. P. Kozak, Y. Zhang and J. R. Sullins are with Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555.

M. A. Shreve and D. B. Goldgof are with Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620.

in [4], "The use of multiple intensity images is of value only if there is some variation between the individual images of a person. And very little is known about how to build the *right* degree of variation into a multi-sample approach". This study is an attempt to address certain aspects of the issue raised above. Although our approach is similar to [4,5], there are significant differences: (1) we use videos of rotating heads that show continuous pose variations; (2) videos have strong shadows which present a severe challenge to the recognition algorithm; (3) fusion is performed on the image level.

A large amount of research efforts have been dedicated to video-based face recognition because of the rich temporal information contained in videos. In the very early work [6], the use of a 3D model has been considered important for both tracking and recognition purposes. Various models have been proposed, from geometrical models to more sophisticated deformable models, morphable models and statistical models [7,8,9,10]. Certain 3D models (meshes, point clouds and depth images) can be directly used for recognition through either registration minimization or principle component analysis. More frequently, 3D models are used to transform an input 2D image by rendering it so that the face in the resulting image has the desired pose, illumination and expression. The drawbacks of using an explicit 3D model are: (1) the accuracy of a reconstructed 3D shape via structure from motion may not be adequate for recognition; (2) the computational cost involved in shape rendering, illumination simulation and deformation modeling is high.

Efforts have been made to extract grey level cues (shading, profile curves and silhouette) to aid the 3D model based recognition [11], because intensity variation is often related to an object's shape and its surface reflectance properties, a fact that has been explored in the well known "shape from shading" scheme. A video sequence of a rotating head should contain abundant information about its 3D geometry that, in theory, can be utilized either explicitly or implicitly [12]. In this paper, our objective is to investigate the feasibility of using multiple video frames (fused on image level) directly for face recognition, without reconstructing a specific 3D model or fitting a generic 3D model.

III. Experiment Design

A. Video Collection

Videos were acquired in two collection sessions, with the second collection being carried out 20 days after the first one. 101 subjects participated in the first collection, among which 47 subjects returned for the second collection (see Table I). The videos of the 47 subjects who enrolled in both collections will be used as gallery and probe sets, while the videos of the remaining 54 subjects who appeared only in the first collection will be used as the training set. Certain subjects showed noticeable changes in their appearance between two sessions, such as beards, mustaches, piercing and glasses (two subjects were allowed to wear eye glasses).

TABLE 1
DATA COLLECTIONS AND LIGHTING CONDITIONS

	First Collection	Second Collection
Subjects	101 subjects.	47 subjects,
Condition One	**Regular indoor light**. Rotation: 90 degrees. Expression: neutral, smile, angry, surprise.	**Regular indoor light**. Rotation: 90 degrees. Expression: neutral, smile, angry, surprise.
Condition Two	**Strong Shadow.** Rotation: 90 degrees. Expression: neutral, smile, angry, surprise.	**Strong Shadow.** Rotation: 90 degrees. Expression: neutral, smile, angry, surprise.

In each collection session, a subject sat on a rotating chair in front of a camcorder against a blue background curtain. The subject slowly turned his/her body by 90 degrees (from the frontal view to the profile view). The turning process was done twice, first with the regular indoor light, and then with strong shadows cast by a headlight. A few samples obtained under the two lighting conditions are shown in Fig. 1.

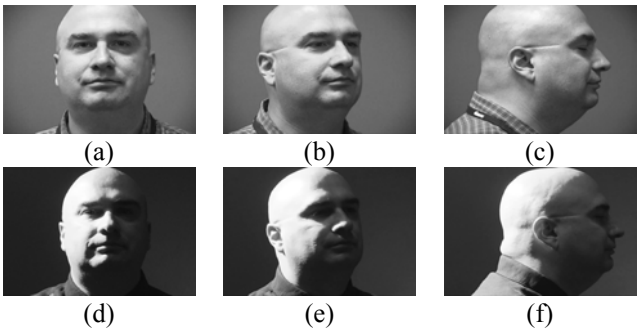


Fig. 1. Sample frames from two video sequences taken under a regular indoor light condition (a, b, c) and a strong shadow condition (d, e, f).

Videos were acquired using a Canon XL1s camcorder with a speed of 30 frames per second. Each rotation resulted in a 10-30 seconds long video sequence, which was then processed with the Adobe Professional software to generate 300 to 900 frames, depending upon the rotation speed. All frames have a resolution of 720 x 480 pixels.

B. Frame Selection

The multi-sample approach requires that a frame pair from the gallery and probe sets must have the same or similar pose angle (rotation degree). Because subjects rotated at different and varying speeds, it is difficult to determine the pose angle in a particular frame accurately. To solve this problem, we developed a software tool that displays the nose positions of user specified angles. As shown in Fig. 2, we chose a coordinate so that the fontal view is 0 degree and the profile view is 90 degrees. X_0 and X_{90} represent the nose positions on X-axis in those two views and are manually marked. Given an arbitrary angle α, its corresponding nose position X_α can be calculated by:

$$X_\alpha = X_0 + (X_{90} - X_0)\sin\alpha$$

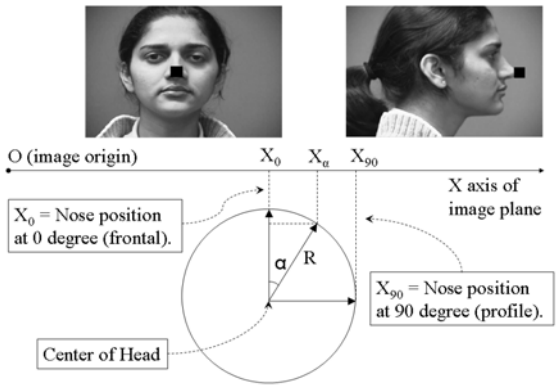
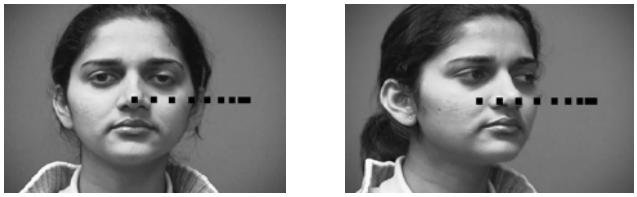


Fig.2. Calculation of the nose position for a particular pose angle. Note that X_0 and X_{90} need to be marked manually.

The software then automatically displays the nose positions of all user specified angles. Fig. 3 illustrates how to use the information to determine that the face has a pose angle of 20 degrees. Note that the nose positions of ten pose angles are displayed (0, 10, 20, 30, 40, 50, 60, 70, 80 and 90).



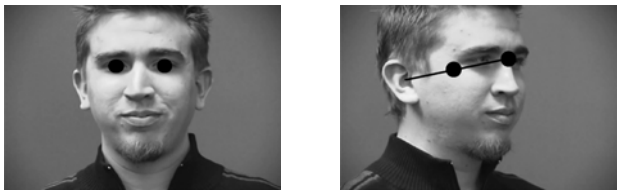
(a) 0 degree rotation (b) 20 degrees rotation

Fig.3. Determination of a frame in which the face has a 20 degrees rotation. The square-shape markers indicate the nose locations of ten pose angles: 0, 10, 20, 30, 40, 50, 60, 70, 80 and 90, from left to right.

C. Fusion on Image Level

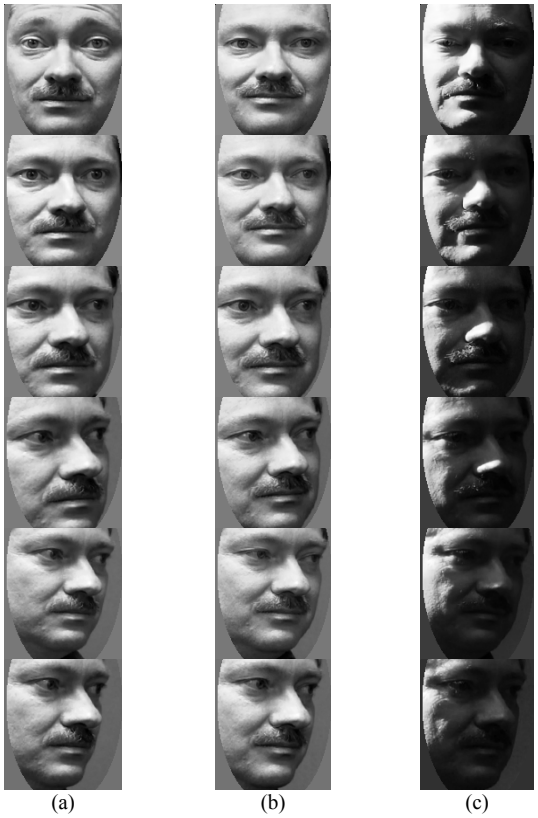
The majority of biometric fusion was done on the score or rank level [13]. Only a few studies have used image (sensor) level fusion. For example, Chang *et al* [14] evaluated the performance of a multi-biometrics system by concatenating a face image and an ear image. Fusion on the image level has the advantage that information in raw data is preserved, and therefore is well suited for the multi-sample approach as long as the number of samples per subject is reasonably small. In this study, we used image level fusion to integrate as many as seven frames per subject.

We performed fusion in three steps: (1) seven frames were chosen for each subject with following rotation degrees: 0, 10, 20, 30, 40, 60 and 90; (2) each frame was normalized using the coordinates of two facial markers. If a face rotated by 0, 10, 20 or 30 degrees, the centers of eyes were used. If a face rotated by 40 degrees, the left corners of eyes were used. If a face rotated by 60 or 90 degrees, the top of the nose and the middle point between the center of the ear and the top of the nose were used (Fig. 4); (3) the normalized images (cropped by an elliptical mask) were then aligned vertically to create a fused image. Fig. 5 shows samples of fused images.



(a) 0 degree rotation (b) 60 degrees rotation

Fig. 4. Selection of two facial markers for image normalization. For 60 degrees rotation, the top of the nose and the middle point between the center of the ear and the top of the nose were used.



(a) Gallery of Test 1, Test 2 (b) Probe of Test 1 (c) Probe of Test 2

Fig. 5. Samples of fused images. (a) Gallery from the first collection under regular light. (b) Probe from the second collection under regular light. (c) Probe from the second collection with shadows.

D. Training, Gallery and Probe Sets

We designed two tests, each with an independent training set. Test-1 includes a gallery of 47 subjects from the first collection under regular light, and a probe of the same 47 subjects from the second collection under regular light. Test-2 consists of the same gallery as Test-1, but with a different probe from the second collection with strong shadows (See Table II and Fig. 5). It is worth noting that the training set of Test-2 must contain frames of both regular light and shadows. Otherwise, the eigenspace will be skewed due to the lack of representative samples, which may cause a large drop in recognition rate. All tests were run with a PCA-based recognition algorithm also known as the "Eigenface" method [15]. Its implementation details can be found in [16].

TABLE II
TRAINING SET, GALLERY SET AND PROBE SET.

	Test 1	Test 2
Training	54 subjects, 378 frames, from the 1st collection, independent from both gallery and probe sets. **Regular indoor light.**	54 subjects, 378 frames, from the 1st collection, independent from both gallery and probe sets. **Regular light + shadow**.
Gallery	47 subjects, 329 frames, from the 1st collection. **Regular indoor light.**	47 subjects, 329 frames, from the 1st collection. **Regular indoor light**.
Probe	47 subjects, 329 frames, from the 2nd collection. **Regular indoor light.**	47 subjects, 329 frames, from the 2nd collection. **Strong shadow.**

IV. RESULTS AND DISCUSSIONS

A. *Test-1: Regular Indoor Light*

The purpose of Test-1 is to examine the performance of multi-sample fusion using different numbers of frames per subject under the regular light. We started with one frame per subject (0 degree), and then fused it with the next frame (10 degrees), until we integrated all seven frames. For example, a 5-frame fusion concatenates frames in an increasing order of rotation degrees: 0, 10, 20, 30 and 40.

The cumulative match characteristic (CMC) curves of Test-1 are plotted in Fig. 6. For visualization purpose, we only show results of using the odd number of frames per subject. An increasing improvement in the rank one recognition rate can be observed, from 91.4% with a single frame, to 95.7% with 3-frame fusion, to 97.8% with 5-frame fusion, and to 100% with both 6-frame fusion and 7-frame fusion. Although the starting rate of the single frame case is relatively high, an almost 10% performance increase is still considered to be significant.

B. *Test-2: Strong Shadow*

One way to assess the robustness and effectiveness of a recognition method is to apply it to images of severe illumination changes. So, we designed Test-2 that has the same data set as Test-1, except that its probe set consists of images from the second collection with shadows on the faces. As can be seen in the sample images (Fig. 1, Fig. 5 and Fig. 8), the shadows almost black out half of the faces. If the method of using fused frames of rotating faces can yield significant performance gain under such an unfavorable condition, its value can be further justified.

Fig. 7 shows the CMC curves of Test-2. As expected, the rank one rate of using a single frame is relatively low (63.8%). But the improvement can be clearly seen as the number of frames used in fusion increases. The rank one rate goes up to 72.3% with 3-frame fusion, 80.8% with 5-frame fusion, and finally to 85.1% with 7-frame fusion, a more than 20% increase. Fig. 8 shows a face that was not recognized until rank 24 using a single frame, but was correctly recognized at rank one with 3-frame fusion.

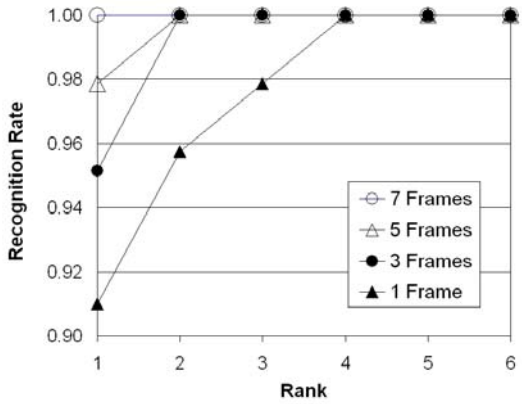


Fig. 6. Performance of multi-sample fusion (on the image level) measured as CMC curves. Both gallery and probe images were taken under the regular indoor lighting condition.

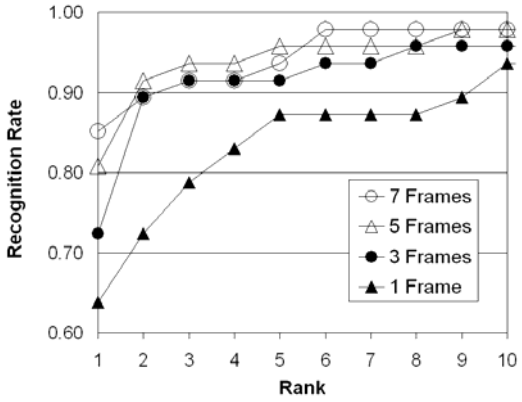


Fig. 7. Performance of multi-sample fusion (on the image level) measured as CMC curves. Gallery images were taken under regular indoor light, while probe images were taken with strong shadows.

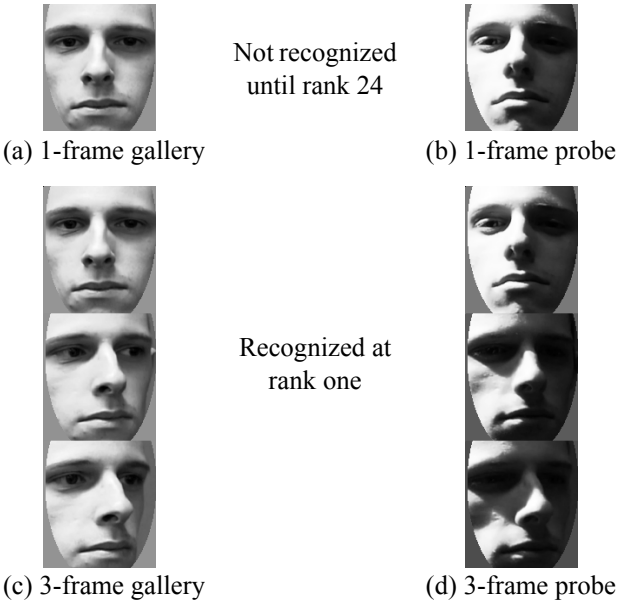


Fig. 8. An example that shows the recognition improvement of using fused images (3-frame) over using a single frame. (a) and (b) were not recognized until rank 24, while (c) and (d) were recognized at rank one.

It should be stressed that not all fusions will result in positive outcomes. Fig. 9 illustrates the rather complex relationship between the rank one recognition rate and the number of frames used in fusion. The overall trend is that the performance improves as the number of frames per subject increases, but their relationship is not strictly monotonic. The recognition rate actually dipped quite a bit in certain cases. For instance, the 2-frame fusion did not improve the performance in both tests. There are several possible explanations: (1) during the rotation, many subjects blinked their eyes because of the headlight; (2) sometimes subjects rotated relatively fast leading to blurred images; (3) there might be a more fundamental issue of multi-sample fusion that is related to the interplay of sample sets and their combined effect. As suggested in [13], if two sets of samples are positively correlated, the noise in the samples could negate any performance gain from their fusion. In the 2-frame fusion case, the performance drop may be explained by the lack of variations between the 0-degree frames and the 10-degree frames. In other words, the faces in those two sets are so similar that their fusion provides little complementary benefit. This explanation seems also consistent with the observation that using multiple identical images achieves the same performance as using one image [3].

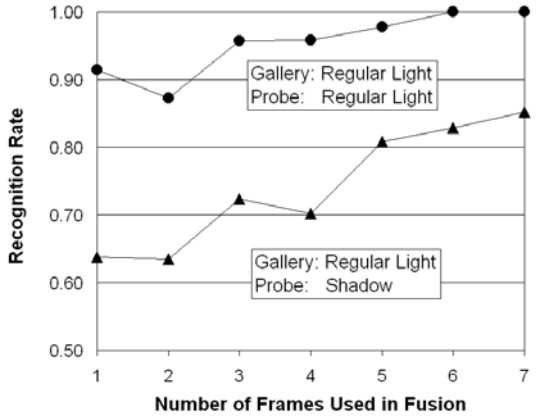


Fig. 9. The relationship between the rank one recognition rate and the number of frames used in fusion.

To gain more insights into the performance of multi-sample fusion from a statistical perspective, we computed the probability distributions for both the match class and the no-match class in Test-2 using the Mahalanobis distance matrices. The match class refers to the gallery-probe pairs from the same person (the diagonal entries in the distance matrices), and the no-match class refers to the gallery-probe pairs from different persons (all non-diagonal entries in the distance matrices). The distributions of using a single frame, 2-frame fusion and 7-frame fusion are shown in Fig. 10, Fig. 11 and Fig. 12, respectively. In the case that a single frame (fontal view) was used, the probability distributions of two classes show large overlaps, suggesting that many samples will be misclassified by a simple minimum distance criterion. On the other hand, the probability distributions of using 7-frame fusion exhibit a much improved separation between the two classes, which explains the observed higher recognition rate. However, there is no noticeable difference between the probability distributions of using a single frame and those of using 2-frame fusion.

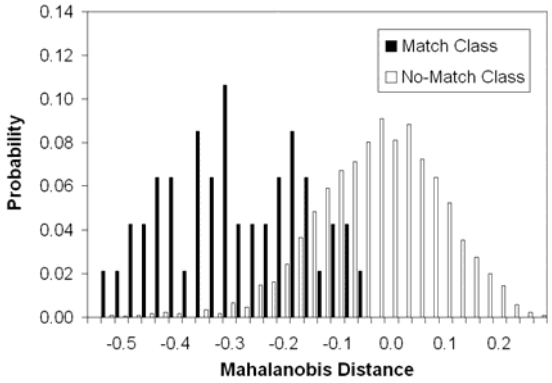


Fig. 10. The probability distributions obtained from the Mahalanobis distance matrices of Test-2 using a single frontal view image.

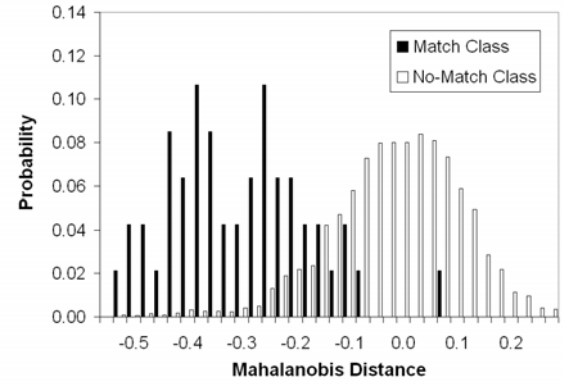


Fig. 11. The probability distributions obtained from the Mahalanobis distance matrices of Test-2 using 2-frame fusion.

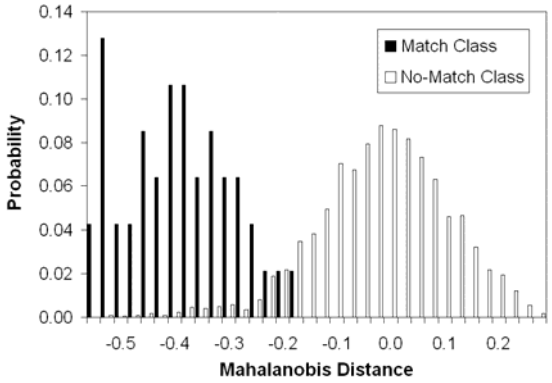


Fig. 12. The probability distributions obtained from the Mahalanobis distance matrices of Test-2 using 7-frame fusion.

V. SUMMARY

To achieve a significant increase in face recognition rate under challenging conditions necessitates the development of new techniques such as 3D scans, high resolution images, multi-sample and multi-modal methods [17]. Using videos that capture the continuous pose and illumination changes of moving faces to provide implicit 3D information is one possible solution. In this paper, we present some preliminary experimental results of recognizing faces that rotated up to 90 degrees using multi-frame fusion. Based on the two tests with videos taken under two lighting conditions, several observations can be made:

1) Recognition rate shows large improvements in both tests, about 10% under the regular lighting condition and about 20% under the strong shadow condition. This performance increase can be, to a large degree, attributed to the coherent intensity variations in video frames that are linked to the 3D geometry of a rotating face and its interaction with lights.
2) A linear function seems inadequate to describe the relationship between the recognition rate and the number of frames used in fusion. It is likely that finding an optimal number of frames to achieve the maximum performance increase will be task-dependent. We will conduct more experiments using 20 to 90 frames per subject with rotation intervals of 1 to 5 degrees. We will also investigate this issue in the framework of scale-space aspect graph so as to find the minimum number of frames that can provide sufficient 3D information for face recognition [18].
3) Fusion of certain frames can lead to performance drops. We present some qualitative analysis based on the probability distributions of two classes. More thorough investigations using the canonical correlation coefficient or a diversity index may shed light on this issue.

Since our motivation is to utilize implicit 3D information in videos via multi-frame fusion, it would be interesting to compare its performance with those of using explicit 3D data such as range images, so that its efficacy can be benchmarked. This requires a data set that includes both range images and rotating head videos of the same subject. We plan to collect those data and double the size of our database to 200 subjects. One related issue is whether image level fusion and score level fusion would yield the same performance, because score level fusion is more computationally attractive if a large number of frames are needed for fusion. Finally, we would like to emphasize that, although a complete video sequence of 90 degrees head rotation is rare in real situations, this kind of data is an ideal testbed that allows us to examine various factors that influence the performance of the multi-sample method. Moreover, in certain realistic scenarios such as video surveillance, even a short video segment that captures partial head rotation could be valuable for recognition.

REFERENCES

[1] K. W. Bowyer, K. Chang, and P. J. Flynn. "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1-15, 2006.

[2] I. A. Kakadiaris, G. Passalis, G. Toderick, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, vol. 29, no. 4, pp. 640-649, 2007.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399-458, 2003.

[4] K. W. Bowyer, K. Chang, P. J. Flynn, and X. Chen, "Face recognition using 2-D, 3-D and Infrared: Is multimodal better than multisample?" *Proceed of the IEEE*, vol. 94, no. 11, pp. 2000-2012, 2006.

[5] D. Thomas, K. W. Bowyer, and P. J. Flynn, "Multi-frame approaches to improve face recognition," *IEEE Workshop on Motion and Video Computing*, pp. 19-19, Austin, TX, 2007.

[6] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705-740, 1995.

[7] V. Blanz, and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063-1074, 2003.

[8] D. DeCarlo, and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *Inter. Jour. of Computer Vision*, vol. 38, no. 2, pp. 99-127, 2000.

[9] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Three-dimensional face recognition," *International Journal of Computer Vision*, vol. 64, no. 1, pp. 5-30, 2005.

[10] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," *Proc. of IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 313-320, 2003.

[11] C. Beumier, and M. Acheroy, "Face verification from 3D and grey-level clues", *Pattern Recognition Letters*, vol. 22, pp. 1321-1329, 2001.

[12] M. Husken, M. Brauckmann, S. Gehlen, K. Okada, C. V. Malsburg, "Evaluation of implicit 3D modeling for pose-invariant face recognition," *Proceedings of SPIE*, vol. 5404, pp. 328-338, 2004.

[13] A. A. Ross, K Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*, Springer, 2006.

[14] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 25, no. 9, pp. 1160-1165, 2003.

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

[16] www.cs.colostate.edu/evalfacerec.

[17] P. J Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Margues, J. Min, and W. Worek, "Overview of the face recognition grand challenge," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2005)*, pp. 947-954, Washington DC, 2005.

[18] D. W. Eggert, K. W. Bowyer, C. R. Dyer, H. I. Christensen, D. B. Goldgof, "The scale space aspect graph", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 15, no. 1, pp. 1114-1130, 1993.