FACE RECOGNITION BY MULTI-FRAME FUSION OF ROTATING HEADS IN VIDEOS

by

Shaun Canavan

Submitted in Partial Fulfillment of the Requirements for the degree of
Master of Science

Department of Computer Science and Information Systems
College of Science, Technology, Engineering and Mathematics

YOUNGSTOWN STATE UNIVERSITY
May, 2008

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

Shaun Canavan

I hereby release this thesis to the public. I understand that this thesis will be made available from the OhioLINK ETD Center and the Maag Library Circulation Desk for public access. I also authorize the University or other individuals to make copies of this thesis as needed for scholarly research.

Signature:

_____

Shaun Canavan, Graduate Student Date

Approvals:

_____

Yong Zhang, PhD, Thesis Co-Major Advisor Date

_____

John Sullins, PhD, Thesis Co-Major Advisor Date

_____

Alina Lazar, PhD, Thesis Committee Member Date

_____

Peter J. Kasvinsky, PhD, Dean of Graduate School Date

ABSTRACT

This paper presents a face recognition study that implicitly utilizes the 3D information in 2D video sequences through a multi-sample fusion process. 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 face models or shapes from range images. The multi-frame fusion was performed on both the image and score levels. Both types of fusion showed large improvements in the recognition rates. The image level fusion showed improvements from 91%, using one frame, to 100%, using 7 frames, under regular lighting. An improvement from 63%, using one frame, to 85%, using 7 frames, was noticed under strong shadow. The score level fusion of two frames also showed an improvement in the recognition rate.

ACKNOWLEDGEMENTS

I would like to thank Dr. Zhang for the many opportunities of doing research with him, especially on the topic of Face Recognition. This research experience allows me to pursue my academics goals to the fullest. I would like to thank Dr. Sullins for being the co-major advisor of my thesis and helping me out on too many occasions to count at YSU. Also, I would like to thank Dr. Lazar for helping me in and out of the class room.

Special thanks go to Mike Kozak, without his help this project would not have been possible. I would also like to thank Cameron and Tracey Hughes, as their feedback and support has been invaluable to me. Finally, to my wife for putting up with so many late nights and weekends that I have spent in the lab.

TABLE OF CONTENTS

Abstract iii

Acknowledgments iv

List of Figures vi

List of Tables vii

1. Introduction 1

2. Related Works 2

3. Experiment Design 4

3.1 Video Collection 4

3.2 Frame Selection 5

3.3 Fusion on Image level 7

3.4 Training, Gallery, and Probe Sets 8

4. Results and Discussions 9

4.1 Test-1: Regular Indoor Light 9

4.2 Test-2: Strong Shadow 10

5. Score Level Fusion 14

6. Conclusions 15

7. References 17

LIST OF FIGURES

Figure 1 **5**

Figure 2 **6**

Figure 3 **7**

Figure 4 **8**

Figure 5 **8**

Figure 6 **11**

Figure 7 **12**

Figure 8 **12**

Figure 9 **13**

Figure 10 **14**

Figure 11 **15**

Figure 12 **16**

LIST OF TABLES

Table 1 **4**

Table 2 **9**

1. Introduction

The field of 3D face recognition has received a great deal of attention as of late, in the biometrics research community. This is due, partially, to the fact that 3D faces are considered to be less affected by the lighting and pose variations that plague 2D faces. When the multi-modal approach is used (2D and 3D combined), there is generally an improvement in the recognition rate. It has been pointed out by *Bowyer et al* [1] and *Kakadiaris et al* [2], that the use of 3D range images has certain limitations. Some of these limitations include:

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.

Most likely these technical issues will not be solved in the near future. Because of this, there is a strong interest to explore other avenues of research to gather 3D information in a non 3D plane.

In this paper, a method is utilized that uses a video sequence in which a subject rotates their head from the 0 degree frontal view to the 90 degree profile view. The hypothesis at hand is that the continuous intensity change of the image stream has the 3D geometry of the face embedded within it. This allows the capture 3D information without the use of an explicit 3D face model. This method has several advantages:

1. If the video sequence, acquired, can provide quality 3D data, some of the constraints of 3D sensors can be relieved. For example, an optical

camcorder can capture the data in real time, which would allow this method to be deployed in real time.

2. The 3D information of the face is implicitly inferred, so the high cost of 3D modeling can be avoided.
3. Not all frames will be used, only a selected number of frames are necessary. This allows that fusion to be performed on both the image and score levels.
4. A video sequence of face with different poses may be able to alleviate some of the adverse effects of lighting.

2. Related Works

For an in-depth discussion of the current developments in 3D methods, this can be found in from *Bowyer et al* [1]. An extensive look at the methods of face recognition is given by *Zhao et al*. [3]. In this section, a review of some of the techniques that are most relevant to this topic will be given.

One motivation for this research is, it has been shown that the multi-sample approach can achieve a performance that is comparable to that of the multi-modal approach. *Bowyer et al*. [4] showed a rank one improvement of 96%, with 2 frames, to 100% with 4 frames. Also, *Thomas et al.* [5] has shown that the recognition rate generally increases as the number of frames increases. They have also found that optimal number of frames to use for this is between 12 and 18. However, [4] has noted that this is only the case if there is sufficient information change in each of the images. At this time very little is known about how to build

the right degree of variation. This research is an attempt to address some of these issues mentioned above. This research is similar to [4, 5], however there are significant differences:

1. Videos in this research show continuous pose variations
2. Strong shadows are used in the videos
3. The fusion is performed on the score and image levels.

There is a great deal of temporal information that is contained in videos; this had lead to a large amount of research in this direction. In the early work [6], the use of a 3D model was considered important for both tracking and recognition purposes. There have been a number of models proposed, from geometrical models to more sophisticated deformable models, morphable models and statistical models [7, 8, 9, and 10]. 3D models have frequently been used to transform a 2D image by rendering so the image has all of the desired changes (light, pose, etc.). Using an explicit 3D model has some drawback:

1. Accuracy of a reconstructed 3D face may not be accurate for recognition, especially those that are built using the structure from motion method.
2. The computational cost of constructing a face model with adequate facial details is extremely high.

Efforts have been made to extract grey level cues to aid the 3D model based recognition [11], because the intensity variation is often related to an object's shape and its surface reflectance properties.

3. Experiment Design

3.1 Video Collections

Two sessions were used to collect data. The second session took place 20 days after the original. 101 subjects participated in the first collection, and 47 of those subjects returned for the second collection (Table 1). The gallery and probe sets consist of the 47 subjects who took part in both collections. The subjects that only took part in the first collection makeup the training set for this research. There were certain subjects that showed noticeable changes in their appearance between the two sessions. Some of these changes include beards, mustaches, piercings and glasses. Normally glasses pose a great problem to face recognition, so for the purpose of experimentation, two subjects were allowed to wear their glasses.

Table 1. Data collection 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.

For each collection session, the subject sat in a rotating chair in front of camera and rotated from the 0 degree frontal view to the 90 degree profile view. This was done against a blue back-drop to reduce background noise. The subject also

performed the previously mentioned rotation twice, once with regular lighting and with a strong shadow for the second rotation. Samples from both of the lighting condition are shown in Fig. 1.

Fig 1. Face images under regular and shadow lighting conditions.

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.

3.2 Frame Selection

When using the multi-sample approach you must be sure that the pair of images from the gallery and probe sets has similar pose angles. It is difficult to determine the actual pose angle due to the subjects' change of speed in rotation. To solve this problem, a software tool was developed that will display the nose positions of user specified angels. As shown in Fig. 2, we chose a coordinate so that the frontal 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:

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

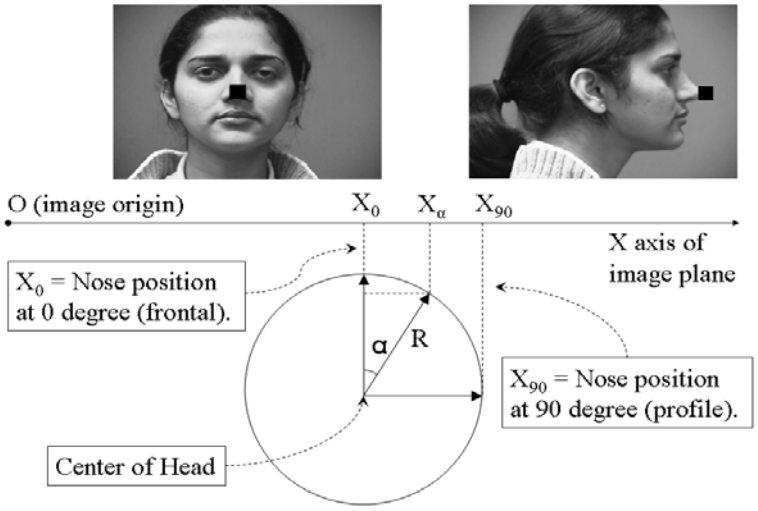

Fig 2. Illustration of determination of pose angles.

The nose positions will then be displayed by the software, depending on which angles the user has specified. In Fig. 3, it is illustrated how the information is used to determine that the selected pose angle is 20 degrees. The nose positions are displayed in ascending order from 0 to 90 degrees in 10 degree increments.

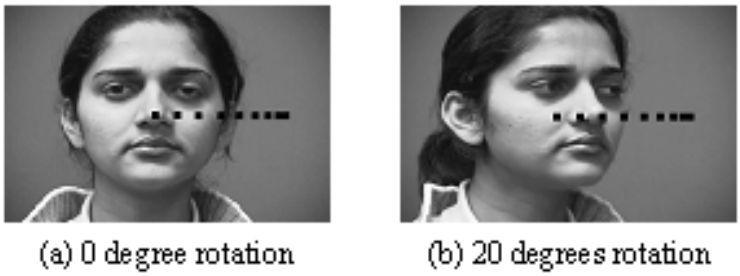

Fig 3. An example that demonstrates the rotation degree selection.

3.3 Fusion on Image Level

In most recent research, most of the fusion was performed on the score level [13]. There have only been a hand-full of studies that have shown the results of image level fusion. *Chang et al.* [14] demonstrated the use of image level fusion by concatenating a face and ear image. By performing the fusion on the image level, one is able to preserve the raw data and thus making for an ample case to study the multi-sample approach. This holds only if the number of images per subject is relatively small.

The fusion was performed in three steps:

1. Seven frames were chosen for each subject with the following rotation degrees: 0, 10, 20, 30, 40, 60 and 90
2. Each frame was normalized using facial markers as shown in Fig. 4
3. The normalized images where then aligned vertically as being shown in Fig.5.

Fig 4. Facial markers used for normalization.

Fig 5. Examples of fused images for both gallery and probe sets.

3.4 Training, Gallery and Probe Sets

Two tests were performed, each with an independent training set. The first test consisted of a gallery with 47 subjects from the first collection, and a probe set with 47, of the same, subjects from the second collection. This was tested under regular lighting conditions only. For the second test, the gallery was the same as the original test, however, the probe set consisted of the subjects under a strong shadow (Table 2). When using the strong shadow for the second test probe set, we must also make sure that the training set contains strong shadow as well. If this is not done the eigenspace will be skewed, due to lack of representative samples. All of the tests that were run used a PCA (Principal Component Analysis) method,

this is also known as the "Eigenface" method [15]. The details of its implementation can be found in [16].

Table 1. Data sets used in the experiments.

Training, Gallery, and Probe Sets		
	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.**

4. Results and Discussions

4.1 Test – 1: Regular Indoor Light

Test one was conducted to examine the performance of multi-sample fusion versus a single frame. The fusion was done in ascending order from top to bottom. The first frame was degree 0, and the last (7^{th}) frame was 90 degrees. This shows key angles from the frontal to the profile view.

The CMC (cumulative match characteristics) curves for test 1 are illustrated in fig. 6. This curve helps to illustrate the recognition rate at ranks one through ten. Only the odd number of frames are shown in this figure, simply for illustration

purposes. Based on figure 6 an improvement from 91%, with one frame, to 100%, with seven frames, can be seen. Although 91% may seem like a high starting point, the recognition rate improved almost 10%. This increase in the recognition rate is significant, especially with extremely large datasets.

4.2 Test – 2: Strong Shadow

Illumination changes offer a severe challenge to face recognition, so one way to test the robustness and effectiveness of this method is to apply it to these changes. Test 2 addresses these challenges. The same dataset is used for test 2, however, the probe consists of images that have a strong shadow cast on them. 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 this method of using fused frames can yield a significant increase in the recognition rate under strong shadows, its value can be further justified.

Fig. 7 illustrates the CMC curves of test 2 under the strong shadows. Since illumination changes offer such a challenge, one can expect the recognition rate to be relatively low compared to that of the regular light. From figure 7 one can see that with only one the frame the recognition rate is only 64%. However, the improvement can clearly be seen as the recognition rate with 7 frames is 84%, an increase of 20%. Fig. 8 illustrates that a single frame image was not recognized until rank 24 (24 rounds of selection), and it was recognized at rank 1 with 3 frame fusion.

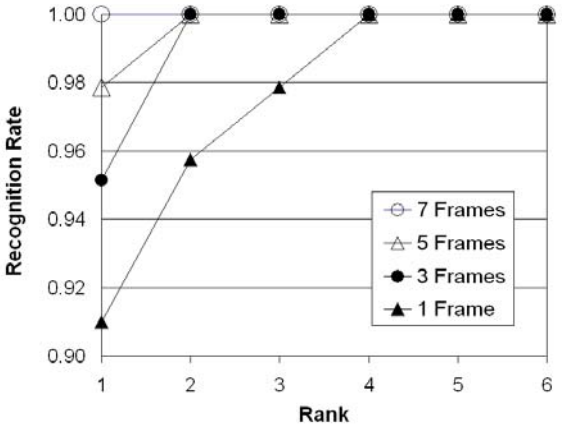

Fig 6. CMC curves of Test-1.

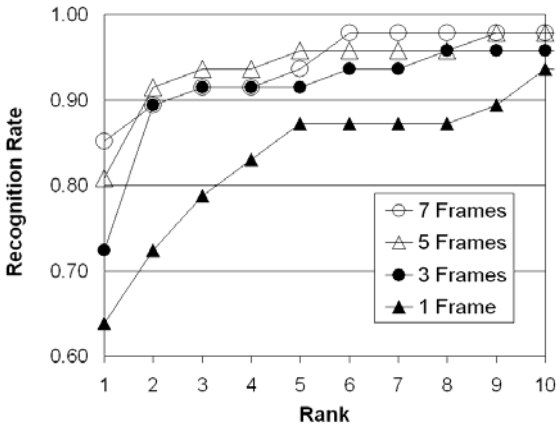

Fig 7. CMC curves of Test-2.

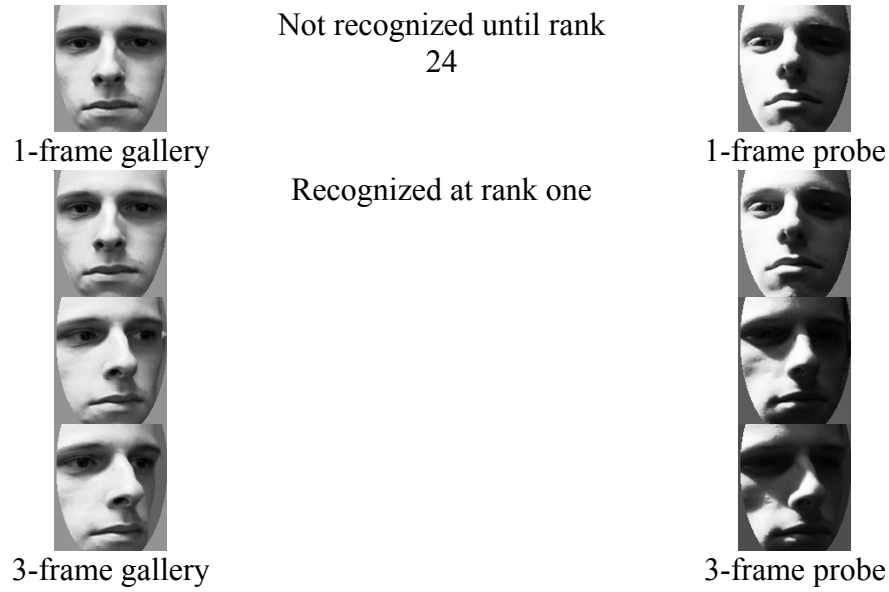

Fig 8. Fusion examples with shadow effects.

It should be stressed that not all fusions will result in positive recognition rates. In Fig. 9, the complex relationship between the rank one recognition rate and the number of frames is illustrated. As can be seen from figure 9, the trend is an improvement in the recognition rate; however, the relationship is not strictly

monotonic. There are several possible explanations for this:

1. During the rotation, some of the subjects blinked their eyes
2. Some of the images were blurred due to the subject moving too fast
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 other words, if two of the images are so similar that the information will be redundant, and add now new additional benefit.

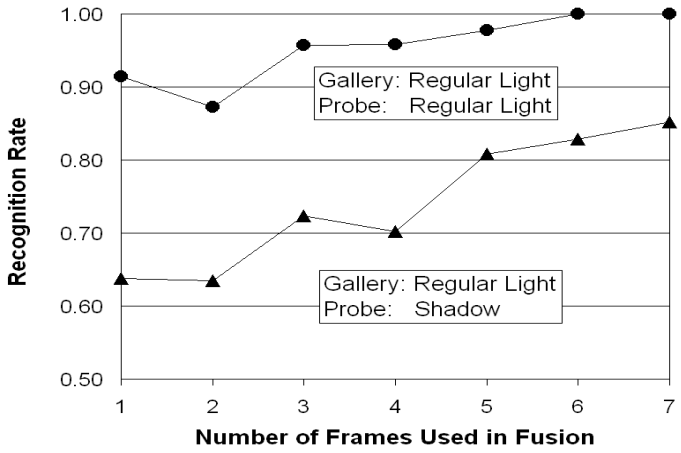

Fig 9. The relationship between rank one rate and number of frames.

To gain more insights into the performance of multi-sample fusion from a statistical view, the probability distributions for both the match and no-match classes were computed (Figs. 10, 11, and 12). The match class refers to the gallery-probes pairs of the same person, and the no-match class refers to the gallery-probe pairs from different people. A single image, 2 frame fusion, and
7 frame fusion are shown respectively.

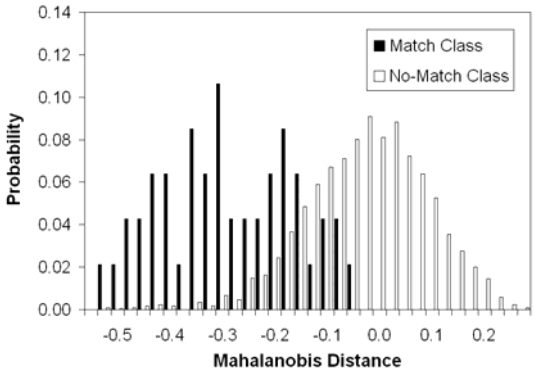

Fig 10. The probability distributions obtained from the Mahalanobis distance matrices of Test-2 using a single frontal view image.

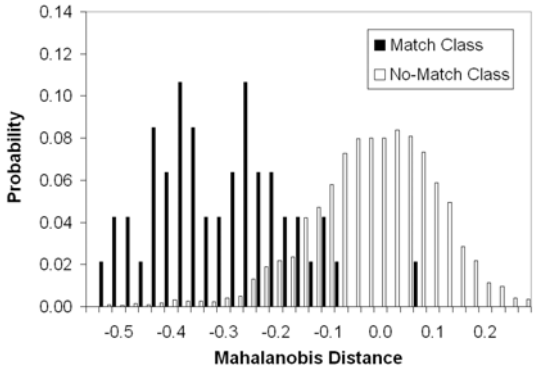

Fig 11. The probability distributions obtained from the Mahalanobis distance matrices of Test-2 using 2-frame fusion.

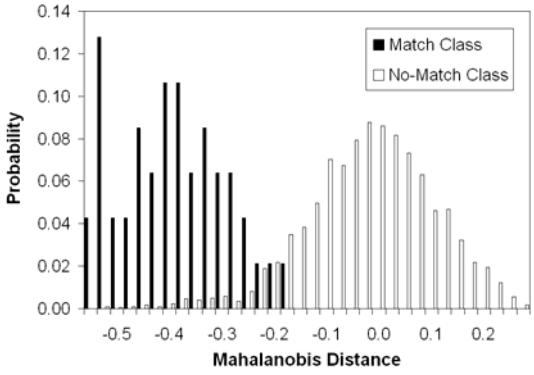

Fig 12. The probability distributions obtained from the Mahalanobis distance matrices of Test-2 using 7-frame fusion.

5. Score Level Fusion

As noted earlier, most of the research involving fusion has been done on the score level; this can be due in part to its simplicity, and relative quickness to perform. To perform the score level fusion the sum rule was used to combine the distance matrixes for each of the images in the dataset. The sum rule was chosen as is has been shown that it is robust to errors in the estimation of the posteriori probabilities [19]. The sum rule can be expressed as follows:

$$Assign\ X \rightarrow w_r\ if$$

$$\sum_{j=1}^{R} (w_r \mid x_j) \geq \sum_{j=1}^{R} P(w_k \mid x_j)$$

it should be noted that this applies when the prior probabilities are equal.

The initial results for score level fusion have been promising, as they have resulted in an increase in recognition rate. The rank one recognition rate of a single frame increased from 91% to 93% when fused with the scores from the 10 degree angle. Although the increase is only 2%, this is significant enough as the fusion only consisted of 2 frames (0 and 10 degrees). It can be hypothesized that adding more frames will lead to rank one recognition rate of 100%. This is based off of the previous work with the image level fusion where the recognition rate did reach 100%. It should also be noted that this is with regular lighting conditions only; the strong shadows have not been researched yet under the score level fusion. The strong shadows are a key component of future research in this direction.

6. Conclusions

New techniques such as 3D scans, high resolution images, and multi-sample methods must be developed to facilitate a significant increase in the face recognition rate [17]. This research, of capturing rotating heads in video under varying lighting conditions is one possible solution to that. Based on the test performed several observations can be made:

1. There was a significant increase in recognition rate for both tests involving image level fusion. An increase of 10% was noted with regular lighting, and an increase of 20% was noted under strong shadows. This increase in performance can be 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. It is not adequate to use a linear function to describe the relationship between the number of frames used and the recognition rate. This can be seen as certain levels of fusion caused a decrease in the recognition rate. Most likely to find the optimal number of frames, it is going to be task-dependent.
3. As noted before fusion of certain frames can lead to a performance drop. Qualitative analyses are given based on the probability distributions of the two classes.
4. The score level fusion also showed an increase in the recognition rate from a single frame to two frame fusion. This is on par with the theory that multi-frame fusion will indeed result in an increased recognition rate.

It would be interesting to compare these results of implicit 3D information to that of explicit 3D information, so that its efficacy can be benchmarked. This would require a dataset that included both range images and rotating heads of the same subject. There is the question as to whether score level fusion and image level fusion would show the same results under the same dataset. Future research in this direction could yield a significant contribution as to whether or not the time required to perform the image level fusion is worth it. It is hard to make a case for this in either direction. Although the recognition rate did increase from one frame to two with the score level fusion, this cannot be used as an accurate indication of how it will perform when more frames are added to the fusion. In theory there would be some information lost in the fusion of the scores, compared to that of the raw data of the image level fusion. This is a very interesting question, and one that should, and will, be investigated further.

6. 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.

[19] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 3, pp. 226-239, Mar. 1998.