Evaluation of Multi-Frame Fusion Based Face Classification Under Shadow

Shaun Canavan[1], Benjamin Johnson[2], Mike Reale[1], Yong Zhang[2], Lijun Yin[1], and John Sullins[2]
State University of New York at Binghamton[1], Youngstown State University[2]

Abstract

A video sequence of a head moving across a large pose angle contains much richer information than a single-view image, and hence has greater potential for identification purposes. This paper explores and evaluates the use of a multi-frame fusion method to improve face recognition in the presence of strong shadow. The dataset includes videos of 257 subjects who rotated their heads by 0° to 90°. Experiments were carried out using ten video frames per subject that were fused on the score level. The primary findings are: (i) A significant performance increase was observed, with the recognition rate being doubled from 40% using a single frame to 80% using ten frames; (ii) The performance of multi-frame fusion is strongly related to its inter-frame variation that measures its information diversity.

1. Introduction

The majority of face recognition researches dealt with still images acquired under a somewhat controlled setting. The performance improvement of recognition technologies using those images has been impressive, as evidenced by the results of Face Recognition Vendor Tests [1]. However, the current methods still have difficulties handling data obtained under more challenging conditions, such as strong shadows, severe occlusions, or large pose variations. To deal with those problems, various approaches have been proposed, including 3D face methods [2], video-based methods [3, 4, 5], correlation-filters [17], multi-view methods [6, 7], and multi-sample/multi-instance methods [8, 9, 10].

In this paper, we examined the performance of a fusion method that integrates multiple frames selected from rotating head videos. The objective was to determine whether and how the multi-frame fusion can overcome the adverse shadow effect to achieve a significantly better recognition rate. We addressed two fundamental issues: (i) How effective is multi-frame fusion in handling shadowed faces, if a sophisticated pre-processing or fusion method (such as a probability density based method) is not involved? (ii) Does a multi-frame fusion yield a consistent performance gain? More importantly, can we quantify its performance in terms of its data composition?

This study has several features: (i) It used a video dataset of 257 subjects, which is comparable to that of Multi-PIE database [11]; (ii) Frames of ten pose angles were automatically selected; (iii) Because of the regular frame interval, the temporal continuity is preserved that characterizes a full head rotation; (iv) A large number of fusion tests were conducted.

2. Related Works

Video-based face recognition bears resemblance to the methods of using multiple still images, but the former may deal with a much larger number of frames. Chellappa *et al*. [4] have developed a probabilistic framework that explores the temporal continuity of face motion. Other approaches of using manifolds and hidden Markov models were also proposed [5, 12]. A probabilistic approach has several advantages: (i) It tackles tracking and recognition simultaneously; (ii) It is flexible to handle both video-to-image and video-to-video matches; (iii) A 3D model can be incorporated. However, the computational cost could be high, especially if a very small frame interval is required to satisfy continuity constraints. Using a high resolution 3D model (e.g., a deformable finite element model) in a video-to-video scenario is even more demanding.

Another popular strategy is to utilize a small number of representative images and consolidate the results through a fusion. Many methods can be put into this category, such as multi-view method, multi-instance method and multi-sample method. Thomas *et al*. [13] and Canavan *et al*. [14] found that recognition rate can be greatly improved using fused video frames. Faltemier *et al*. [8] applied a similar strategy to a 3D face dataset and found that the multi-instance method outperforms a component based method. Recently, a mosaicing approach was proposed that utilizes a composite model from images of different poses [6].

3. Multi-Frame Fusion Based Method

3.1. Video Dataset

Videos of 257 subjects were collected in two sessions. The second session occurred about 5-9 weeks after the first one. 167 subjects attended both sessions and 90 subjects appeared in the first session only. During each session, subjects rotated their heads in the range of 0° to 90°. Two illumination conditions were

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.315

considered: (i) Normal indoor lighting; (ii) Strong shadow. Figure 1 shows some examples.

Figure 1. Upper four rows: Sample frames showing the different views with two different illuminations. Bottom row shows an example of two-sessions: Left four: first session with two different illuminations; Right two: second session of the same subject.

3.2. Frame Selection

Ten frames were selected from each video corresponding to ten pose angles (0°, 10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°, 90°), with 0° for the frontal view and 90° for the profile view. Both manual and automatic methods were used. Manually selected frames were used to benchmark the automatically selected ones. We applied a PCA approach for automatic pose estimation. We collected training data from BU-3DFE database [18] with ten different views from 0° to 90°. After applying the PCA transformation, we obtain the eigen-faces with different views. In the eigen-space, ten clusters are clustered corresponding to ten poses. Given a face image, we project it to the eigen-space and classify to one of the cluster using a K-NN classifier. Following this procedure, we detect ten poses from the video input (see [19] for details). The automatic pose detection process allows us to study the multiple-pose fusion performance in the subsequent experiment.

3.3. Training, Gallery and Probe Sets

The training set contains 90 subjects who appeared only in the first session. The gallery/probe sets include 167 subjects who enrolled in both sessions. The gallery has the frames of normal lighting condition, while the probe has frame of shadows (Table 1). This protocol is similar to that of FRVT 2006 [1], which ensures the independence between the training and test data.

Table 1. Training, Gallery and Probe Sets.

Training	Gallery	Probe
90 subjects. In the 1st session **only**. Normal + Shadow.	167 subjects. In the 1st session. **Normal lighting.**	167 subjects. In the 2nd session. **Strong shadow.**

It should be emphasized that, besides shadows, a few other factors make the dataset very challenging. As shown in Figure 2, there exit large discrepancies between the appearances of the same person in gallery and probe, which could be caused by facial expressions, glasses, jewelry, mustaches and long hair.

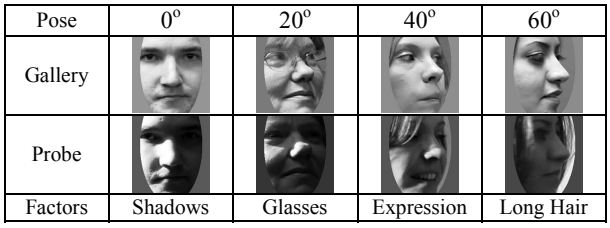

Figure 2. Large differences between faces in the gallery and probe sets that could cause problems to the methods that use a single image per subject.

3.4. Fusion Schemes

Each of the facial poses provides a matching score, which is a similarity measure (e.g., distances) between the images. We used a score level fusion method [9] that was implemented in two steps. In the first step, ten basic score matrices were generated using a PCA (Principle Component Analysis) eigen-face method [15, 16], one for each of the ten pose angles. For example, to create a basic score matrix for the 20° pose angle, a PCA test would be run using only the frames of 20° in the training, gallery and probe sets. In the second step, fusions were carried out by combining the subsets of ten basic matrices with the *sum rule* [9, 14] (i.e., summation of the scores.) Therefore, an exhaustive evaluation requires a total of 1023 fusion tests: $1023 = \boldsymbol{C}(10, 1) + \boldsymbol{C}(10, 2) + \ldots + \boldsymbol{C}(10, 10)$, where $\boldsymbol{C}(n, k) = n!/(k!(n-k)!)$ is the binomial coefficient (see Table 2).

Table 2. Exhaustive Fusion Tests.

Fusion Group	Examples of frame combinations
C(10, 1) = 10	(0°), (10°), (20°), (30°), (40°), (50°), (60°), (70°), (80°), (90 °)
C(10, 2) = 45	(0°, 10°), (80°, 90°)
C(10, 3) = 120	(0°, 10°, 20°), (40°, 80°, 90°)
C(10, 4) = 210	(0°, 10°, 20°, 30°), (10°, 20°, 60°, 90°)
C(10, 5) = 252	(0°, 10°, 20°, 30°, 40°), (10°, 30°, 40°, 60°, 80°)
C(10, 6) = 210	(0°,10°,20°,30°,40°,50°)
C(10, 7) = 120	(0°,10°,20°,30°,40°,50°,60°)
C(10, 8) = 45	(0°,10°,20°,30°,40°,50°,60°,70°)
C(10, 9) = 10	(0°,10°,20°,30°,40°,50°,60°,70°,80°)
C(10, 10) = 1	(0°, 10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°, 90°)
Total = 1023	

3.5. Measuring Inter-frame Variation

In order for a multi-frame method to be effective, the frames used in a fusion should be as diverse as possible (i.e., smaller similarity). So, we adopted a similarity measure based on the mutual information. For two frames A, B, let $P_A(a)$ be the probability density that a

point chosen (uniformly) at random is of intensity a in frame A, and let $P_{A,B}(a,b)$ be the joint probability density that a point chosen at random is of intensity a in frame A, and the same point is of intensity b in frame B. Then the similarity measure $I(A,B)$ is defined as follows:

$$I(A,B) = \iint P_{A,B}(a,b) \log\left(\frac{P_{A,B}(a,b)}{P_A(a)P_B(b)}\right) dadb \quad (1)$$

Using $I(A,B)$, we devised an inter-frame variation metric for a 2-frame fusion:

$$\tau_2(i,j) = \frac{\sum_{k=1}^{N}\left(\frac{1}{I_k(i,j)}\right)}{N}, \quad i,j \in [0^0, 10^0, \ldots, 90^0], \; i \neq j. \quad (2)$$

where τ_2 denotes inter-frame variation, N is the size of a data set. In other words, τ_2 measures the dissimilarity of two frames averaged over all subjects in a data set. In case that a fusion has more than two frames, we first calculate the τ_2 values of all possible 2-frame pairs and then take their average as the τ of that fusion.

Table 3. Statistics of Rank-1 Rate of Fusion Tests.

Fusion Group	Rank-1 Rate			
	Min	Max	Average	Std. Dev.
C(10, 1) = 10	0.31	0.48	0.39	0.05
C(10, 2) = 45	0.41	0.62	0.53	0.05
C(10, 3) = 120	0.50	0.71	0.62	0.05
C(10, 4) = 210	0.56	0.78	0.67	0.04
C(10, 5) = 252	0.59	0.81	0.71	0.04
C(10, 6) = 210	0.66	0.81	0.74	0.03
C(10, 7) = 120	0.70	0.81	0.76	0.03
C(10, 8) = 45	0.73	0.81	0.77	0.02
C(10, 9) = 10	0.75	0.81	0.78	0.02
C(10, 10) = 1	0.78	0.78	0.78	N/A

4. Experimental Results and Discussions

4.1. Improvement by Multi-frame Fusion

The rank-1 rates of 1023 fusion tests were summarized in Table 3, and were plotted in Figure 3 and Figure 4 for CMC curves of a fusion test series. It is clear that the performance of multi-frame fusion steadily improves as the number of frames increases. On average, the fusion method almost doubled the recognition rate, from 40% with a single frame to 80% with ten frames. This is a significant improvement, considering that the dataset used is quite challenging.

In a fusion group that has the same number of frames, the recognition rate showed some fluctuations. For example, in the 3-frame group, the fusion of (0^o, 40^o, 90^o) had the highest recognition rate of 0.713, while the fusion of (70^o, 80^o, 90^o) had the lowest value of 0.503. However, as the number of frames in a fusion increased, the differences among individual fusion tests became less noticeable. At the same time, the fusion performance also leveled off. Adding more frames would not lead to a sizable performance gain. This saturation effect was also observed in other studies [13, 14], suggesting the existence of a performance upper-bound that is likely dependent upon the quality of dataset being used as well as the efficiency of recognition and fusion algorithms.

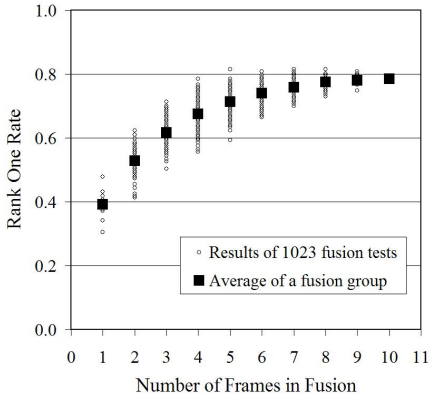

Figure 3. Relationship between the rank-1 rate and the number of frames used in fusion. For each group of fusion tests that contains the same number of frames, its average was also shown.

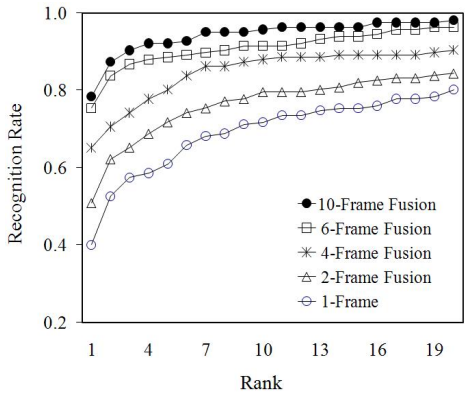

Figure 4. The CMC curves of a fusion test series: (0^o), (0^o, 10^o), (0^o, 10^o, 20^o), …, (0^o, 10^o, 20^o, …, 80^o, 90^o). For visualization purpose, only 1-frame, 2-frame, 4-frame, 6-frame, and 10-frame tests are shown.

4.2. Inter-frame Variation

Since a fusion group of the same number of frames but different combinations showed large recognition rate variations, it is important to seek the underlying cause in a quantitative fashion. To this end, we calculated an inter-frame variation value for each fusion test using Eq. (2). The results of three representative groups (2-frame, 3-frame and 5-frame) were plotted against the Fusion Improvement Ratio (FIR) in Figure 5. The FIR was computed by:

$$FIR = R_m / [(\sum_{i=1}^{m} r_i) / m], \qquad i \in m \quad (3)$$

where R_m is the recognition rate of an m-frame fusion, r_i is the single-frame recognition rate using the ith member of the m frames. So, FIR measures the

performance improvement of an *m*-frame fusion over the average of its individual members.

A positive correlation between the FIR and the inter-frame variation can be observed (Figure 5). This suggests that a fusion of more diverse samples is likely to produce a better recognition rate. For example, in a 4-frame group, the fusion of (0^o, 20^o, 40^o, 90^o) had the highest recognition rate of 0.784, while the fusion of (60^o, 70^o, 80^o, 90^o) gave the lowest rate of 0.557. Apparently, (0^o, 20^o, 40^o, 90^o) is more representative of a full 90 degree head rotation than (60^o, 70^o, 80^o, 90^o) is, because the faces in 60^o, 70^o, 80^o, and 90^o poses are very similar to each other (see Figure 1). In other words, the first fusion combination reveals more about the 3D shape of a face than the second one does.

The above observations is also extendible to the multi-sample approach, multi-enrollment approach, and even the multi-modal approach, where the selection of samples or biometric modalities should be guided by certain inter-sample or inter-modality variation index in order to maximize the performance gain.

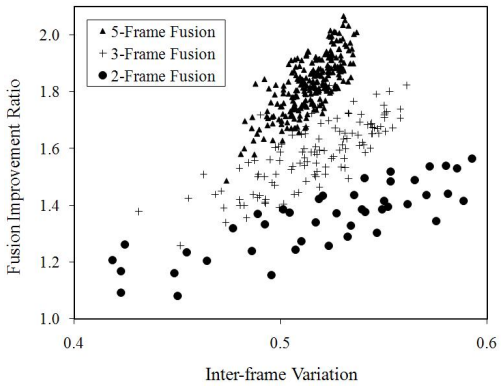

Figure 5. Relationship between the inter-frame variation and the FIR (Fusion Improvement Ratio).

5. Conclusions

This paper presents a multi-frame fusion study and evaluation that exploits the coherent intensity variations in head rotation videos to facilitate recognition under adverse shadow conditions. Based on the tests of 1023 fusion combinations using 257 subjects and 10 frames per subject, following observations can be made: (i) multi-frame fusion is an effective method to improve video face recognition. In a multi-frame to multi-frame scenario, the recognition rate was almost doubled; (ii) the performance of a particular fusion choice has a strong connection to its inter-fame variation. Our future work will verify it on Multi-PIE [11] and investigate the weighted average approach and image-level based fusion for multi-view and multi-frame face recognition.

6. Acknowledgements

This work was supported in part by the URC grant 08-#8 at YSU and NYSTAR James Investigator Program and the NSF under grant IIS-0414029 at Binghamton University.

7. References

[1] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. "FRVT 2006 and ICE 2006 large-scale results", *National Inst. of Standards and Tech., Internal Report* 7408, 2007.

[2] K. W. Bowyer, K. Chang, and P. J. Flynn. "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *CVIU,* 101(1):1-15, 2006.

[3] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video", *Computer Vision and Image Understanding*, 91, pp. 214-245, 2003.

[4] R. Chellappa and S. Zhou, "Face tracking and recognition from video", *Handbook of Face Recognition*, S. Li and A. K. Jain (Eds.), Springer, 2004.

[5] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," *IEEE CVPR*, pp. 313-320, 2003.

[6] R. Singh, M. Vatsa, A. Ross, and A. Noore, "A mosaicing scheme for pose invariant face recognition", *IEEE Trans. on SMC (B)*, 37(5):1212-1225, 2007.

[7] R. Gross, S. Baker, I. Matthews, and T. Kanade, "Face recognition across pose and illumination", *Handbook of Face Recognition*, S. Li and A. K. Jain (Eds.), Springer, 2004.

[8] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "Using multi-instance enrollment to improve performance 3D face recognition", *CVIU*, 112(2):114-125, 2008.

[9] T. Kim and J. Kittler, "Design and fusion of pose-invariant face-identification experts", *IEEE Trans. on Circuits & Sys. for Video Tech.*, 16(9), pp. 1096-1106, 2006.

[10] Y. Zhang and A. Martinez, A weighted probabilistic approach to face recognition from multiple images and video sequences, *Image & Vision Computing* 24(6):626-638, 2006.

[11] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE", *IEEE FGR*, 2008.

[12] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden Markov models", *IEEE CVPR*, 2003.

[13] D. Thomas, K. W. Bowyer, and P. J. Flynn, "Strategies for improving face recognition from video", *Advances in Biometrics: Sensors, Systems and Algorithms*, N. Ratha and V. Govindaraju, editors, Springer, 2007.

[14] S. Canavan, M. Kozak, Y. Zhang, S. Sullins, M. Shreve, and D. Goldgof, "Face recognition by multi-frame fusion of rotating heads in videos", *IEEE Conf. on BTAS 2007*.

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

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

[17] M. Savvides, B. Vijaya Kumar, and P. Khosla, "'Corefaces'- Robust Shift Invariant PCA based correlation filter for illumination tolerant face recognition," In *IEEE CVPR 2004* pp. 834-841.

[18] L. Yin, et al, A 3D facial expression database for facial behavior research, *IEEE FGR*, 2006.

[19] M. Reale, Head pose determination, tracking, and gaze estimation for HCI, *Master thesis*, Binghamton Univ., 2009.