Appeared in Proc. of IEEE Computer Society Workshop on Biometrics (CVPRW), (Minneapolis, USA), June 2007

Face Recognition in Video: Adaptive Fusion of Multiple Matchers

Unsang Park and Anil K. Jain
Michigan State University
East Lansing, MI 48824, USA
{parkunsa,jain}@cse.msu.edu

Arun Ross
West Virginia University
Morgantown, WV 26506, USA
arun.ross@mail.wvu.edu

Abstract

Face recognition in video is being actively studied as a covert method of human identification in surveillance systems. Identifying human faces in video is a difficult problem due to the presence of large variations in facial pose and lighting, and poor image resolution. However, by taking advantage of the diversity of the information contained in video, the performance of a face recognition system can be enhanced. In this work we explore (a) the adaptive use of multiple face matchers in order to enhance the performance of face recognition in video, and (b) the possibility of appropriately populating the database (gallery) in order to succinctly capture intra class variations. To extract the dynamic information in video, the facial poses in various frames are explicitly estimated using Active Appearance Model (AAM) and a Factorization based 3D face reconstruction technique. We also estimate the motion blur using Discrete Cosine Transformation (DCT). Our experimental results on 204 subjects in CMU's Face-In-Action (FIA) database show that the proposed recognition method provides consistent improvements in the matching performance using three different face matchers.

1. Introduction

Automatic face recognition has been actively studied for over three decades as a means of human identification. While substantial improvements in recognition performance have been made under frontal pose and optimal lighting conditions, the recognition performance severely degrades with pose and lighting variations [1, 2, 3]. Therefore, most of the current research related to face recognition is focused on addressing the variations due to pose and ambient lighting.

While conventional face recognition systems mostly rely upon still shot images, there is a significant interest to develop robust face recognition systems that will take advantage of video and 3D face models. Face recognition in video, in particular, has gained large attention due to the widespread deployment of surveillance cameras. Ability to automatically recognize faces in real time from video will facilitate the covert method of human identification using existing network of surveillance cameras. However, face images in video often contain non-frontal poses of the face and undergo severe lighting changes, thereby impacting the performance of most commercial face recognition systems. Even though there are other important problems in video surveillance such as human tracking and activity recognition [18, 19], we will limit our focus only to the face recognition problem in this paper.

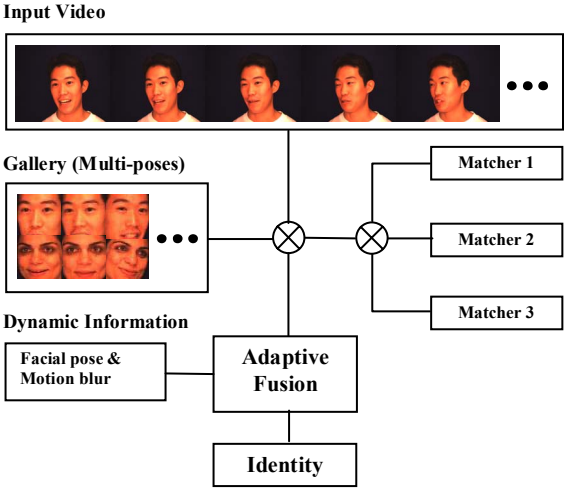


Figure 1: Schematic of the proposed face recognition system in video.

Two well known approaches to overcome the problem of pose and lighting variations are view-based and view synthesis methods. View-based methods enroll multiple face images under various pose and lighting conditions, and match the probe image with that gallery image which has the most similar pose and lighting conditions [4, 5]. View-synthesis methods, on the other hand, generate synthetic views from the input probe image that have similar pose and lighting conditions as the gallery images in order to improve the matching performance. The desired view can be synthesized by learning the mapping function between pairs of training images [6] or by using 3D face models [7, 16]. The parameters of the 3D face

Table 1: A comparison of video based face recognition methods.

	Approaches	No. of Subjects	Accuracy
Chowdhury et al. [20]	Frame level matching with synthesized gallery from 3D model	32	90%
Lee et al .[23]	Matching frames with appearance manifolds obtained from video	20	92.1%
Zhou et al. [21]	Frame to video and video to video matching using statistical models	25 (video to video)	88~100%
Liu et al. [22]	Video level matching using HMM	24	99.8%
Aggarwal et al. [24]	Video level matching using autoregressive and moving average model (ARMA)	45	90%
Proposed method	Frame level matching using fusion of multiple matchers with dynamic information of video	204	Up to 99%

model in the view synthesis process can also be used for face recognition [7].

The view-synthesis approach has the following two advantages over the view-based method: i) it does not require the tedious process of collecting multiple face images under various pose and lighting conditions, and ii) it can generate synthetic frontal facial images under favorable lighting conditions on which state-of-the-art face recognition systems can perform very well. However, some of the disadvantages of the view-synthesis approach are: i) the need for a complicated training process, ii) pose or lighting correction process can introduce noise that may further degrade the original face image, and iii) fragile synthesis process.

There also have been a number of studies that perform face recognition specifically on video streams. Chowdhury et al. [20] estimate the pose and lighting of face images contained in video frames and compare them against synthetic 3D face models exhibiting similar pose and lighting. However, the 3D face models are registered manually with the face image in the video. Lee et al. [23] propose an appearance manifold based approach where each database or gallery image is matched against the appearance manifold obtained from the video. The manifolds are obtained from each sequence of pose variations. Zhou et al. [21] propose to obtain statistical models from video using low level features (e.g., by PCA) contained in sample images. The matching is performed between a single frame and the video or between two video streams using the statistical models. Liu et al. [22] and Aggarwal et al. [24] use HMM and ARMA models, respectively, for direct video level matching. Most of these video based approaches provide good performance on small databases, but need to be evaluated on a larger database. Table 1 summarizes some of the major video based recognition methods discussed in the literature.

We propose a view based face recognition system using video that improves the performance by dynamically fusing the matching results from multiple frames and multiple matchers. The dynamic information is represented as multiple facial poses and motion blur present in the video. The illumination variation is not considered in this work. The proposed system is depicted in Fig. 1. While the static fusion is commonly used in the literature, the adaptive fusion is more suitable for face recognition in video since it actively accumulates the identity evidence according to the dynamic nature of the data. We use the AAM [11] and 3D shape reconstruction [15] process to accurately estimate the facial pose in each frame. The three facial matchers considered in this work are FaceVACS from Cognitec [8], a PCA technique [12] and a cross correlation matcher [10]. These PCA and cross correlation matchers are commonly used as baseline face matchers.

The contributions of our work are i) using multiple face matchers to complement the matcher performance depending on the facial characteristics, ii) designing an adaptive fusion technique for multiple matchers across multiple poses and motion blur and iii) demonstrating the dependency of the recognition performance on the correlation between gallery and probe data. We evaluate the proposed method on a large public domain video database, FIA, [9] containing 221 different subjects (204 subjects were used in our experiments).

2. Problem Statement

The problem is to determine a subject's identity in a video based on the matching scores obtained from multiple face matchers across multiple frames. Consider a video stream with r frames and assume that the individual frames have been processed in order to extract the facial objects present in them. Let $t_1, t_2, \ldots, t_r$ be the feature sets computed from the faces localized in the r frames. Further, let $w_1, w_2, \ldots, w_n$ are the n identities enrolled in the authentication system and $g_1, g_2, \ldots g_n$, respectively, be the corresponding feature templates associated with these identities. The first goal is to determine the identity of the face present in the i^{th} frame as assessed by the k^{th} matcher. This can be accomplished by comparing the extracted feature set with all the templates in the database in order to determine the best match and the associated identity. Thus,

$$ID_i = \arg \max_{j=1,2,\ldots n} S_k(t_i, g_j), \quad (1)$$

where ID_i is the identity in the i^{th} frame and $S_k(\cdot)$ represents the similarity function employed by the k^{th} matcher to compute the match score between feature sets t_i and g_j. If there are m matchers, then a fusion rule may be employed to consolidate the m match scores. While there are several fusion rules, we employ the simple sum rule (with min-max normalization) to consolidate the match scores, i.e.,

$$ID_i = \arg \max_{j=1,2,\ldots n} \sum_{k=1}^{m} S_k(t_i, g_j). \quad (2)$$

Now the identity of a subject in the given video stream can be obtained by accumulating the evidence across the r frames. In this work, the maximum rule is employed to facilitate this, i.e., the identity that exhibits the highest match score in the r frames is deemed to be the final identity. Therefore,

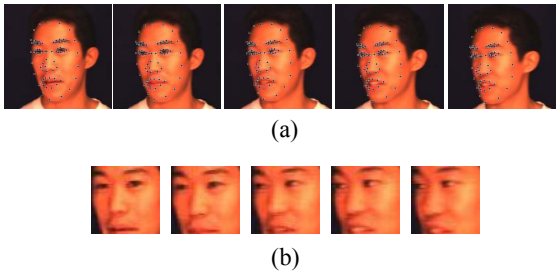
$$ID = \arg \max_{j=1,2,\ldots,n} \left(\arg \max_{i=1,2,\ldots,r} \left(\sum_{k=1}^{m} S_k(t_i, g_j) \right) \right). \quad (3)$$

In the above formulation, it must be noted that the feature sets t_i and g_j are impacted by several different factors such as facial pose, ambient lighting, motion blur, etc. If the vector θ denotes a compilation of these factors, then the feature sets are dependent on this vector, i.e., $t_i \approx t_i(\theta)$ and $g_j \approx g_j(\theta)$. In this work, $m = 3$ since three different face matchers have been used and the vector θ represents facial pose and motion blur in video. The dynamic nature of the fusion rule is explained in subsequent sections.

3. Face Recognition Engines and Database

We use one off-the-shelf face matcher, FaceVACS from Cognitec, and two generic face matchers. The FaceVACS, which performed very well in the FRVT 2002 and FRVT 2006 competitions [2, 26], is known to use a variation of Principle Component Analysis (PCA) technique. However, this matcher has limited operating range in terms of facial pose. To overcome this limitation and to facilitate continuous decisions on the subject's identity across multiple poses, the conventional PCA [12] based matcher and cross correlation based matcher [10] were also considered. The PCA engine calculates the similarity between probe and gallery images after applying the Karhunen-Loeve transformation to both the probe and galley images. The cross correlation based matcher calculates normalized cross correlations between the probe and gallery images to obtain the matching score.

We use CMU's Face In Action database [9] collected in three different indoor sessions and another three different outdoor sessions. The number of subjects varies across the different sessions. We use the first indoor session in our experiments because it i) has the largest number of subjects (221), ii) contains significant number of both frontal and non-frontal poses and iii) has negligible lighting variations. Each video of a subject consists of 600 frames. We partition the video data into two halves and use the first half as gallery data and the second half as probe data.



(a) (b)

Figure 2: Example of cropping face images based on the feature points. (a) Face images with AAM feature points and (b) corresponding cropped face images.

4. Tracking Feature Points

Facial pose is an important factor to be considered in video based face recognition. We detect and track a set of facial feature points and estimate the facial pose based on these feature points. The Active Appearance Model (AAM) has been used to detect and track facial feature points. We use the Viola-Jones face detector [25] and reject the detected points when they deviate substantially from the estimated face area. The AAM feature points are also used to tightly crop the face area to be used by the PCA and cross correlation matchers. Figure 2 shows examples of AAM-based feature tracking and the resulting cropped face images.

4.1. Active Appearance Model (AAM)

AAM is a statistical appearance model generated by combining shape and texture variation [11]. AAM requires a set of training data with annotations $x_1, x_2, \ldots, x_n$ where x_i represents a set of points marked on the image i. Exact correspondences are required in x across all training images. By applying PCA to x, any x_i can be approximated by

$$x_i = x_\mu + P_s \cdot b_s, \quad (4)$$

where x_μ is the mean shape, P_s is a set of orthogonal modes of variation obtained by applying PCA to x and b_s is a set of shape parameters. To build a texture model,

each example image is warped so that its control points match the mean shape. Then the color value g is obtained by the region covered by the mean shape. The texture model is defined similarly with shape model as

$$g_i = g_\mu + P_g \cdot b_g, \tag{5}$$

where g_μ is the mean texture, P_g is a set of orthogonal modes of variation obtained by applying PCA to g and b_g is a set of texture parameters. Once b_s and b_g are obtained, any new shape and associated texture can be approximated by $b = (b_s, b_g)$. Now the problem becomes one of finding the best b that achieves the minimum difference between the test image I_i and the image I_m generated by the current model defined by b. Details about an efficient way of searching the best model parameter b can be found in [11].

4.2. AAM Training

Instead of using a single AAM for multiple poses, we use multiple AAMs, each for a different range of poses [17]. In this way each model is expected to find better facial feature points for its designated pose. Moreover, the number of feature points in each AAM can be different according to the pose (e.g., frontal vs. profile). We chose seven different AAMs for frontal, left half profile, left profile, right half profile, right profile, lower profile and upper profile to cover the pose variations appearing in video data. Assuming facial symmetry, the right half and right profile models are obtained from the left half and left profile models, respectively.

The off-line manual labeling of feature points for each training face image is a time consuming task. Therefore, we use a semi-automatic training process to build the AAMs. The training commences with about 5% of the training data that has been manually labeled, and the AAM search process is initiated for the unlabelled data. Training faces with robust feature points are included into the AAM after manually adjusting the points, if necessary. The AAM facial feature searching process is then initiated again. This process is repeated until all the images in the training set have been labeled with feature points. Our proposed scheme uses a generic AAM where the test subject is not included in the trained AAM. To simulate this scenario we generate two sets of AAMs to simulate the role of training and test images.

4.3. Facial Pose Estimation

The 3D shape model based on the AAM points is reconstructed using the Factorization algorithm [15]. The facial pose can be calculated directly using the factorization process when the pose information of a single frame is available as a reference. However, the direct solution exhibits large errors when there is measurement noise in facial feature point detection. Moreover, the direct solution cannot be obtained in cases where the factorization fails. We will briefly review the Factorization algorithm and show where it fails. Let W, M and S respectively represent the 2D facial feature points, rotation matrix and 3D shape, then the Factorization is formulated as

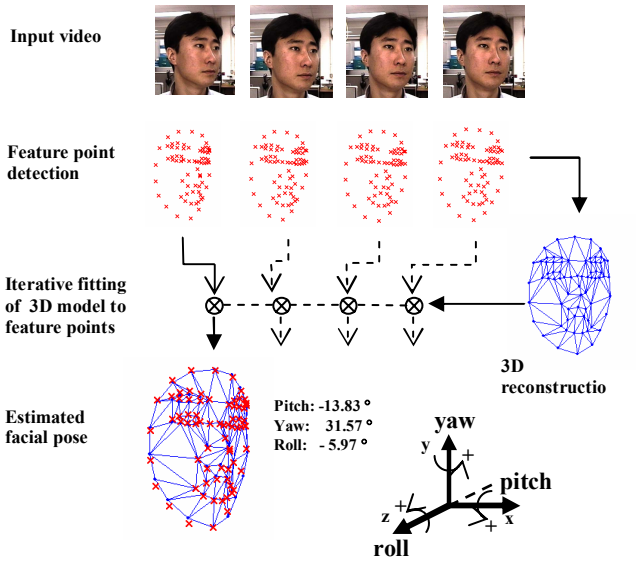


Figure 3: Schematic of facial pose estimation.

$$W = M \cdot S. \tag{6}$$

The translation term is omitted in Eq. (6) because it is already compensated for in W. By applying singular value decomposition, W is factored as $U \cdot D \cdot V^T$. The U, D and V^T matrices initially have dimensions of $2F \times 2F$, $2F \times P$ and $P \times P$, respectively, where F is the number of frames and P is the number of facial feature points. Then, the size of U, D and V are reduced to U', D' and $V^{T'}$ with dimensions of $2F \times 3$, 3×3 and $3 \times P$, respectively, according to the top three singular values to meet the rank-3 constraint. The initial estimate of M and S, i.e., M' and S', become $U' \cdot D'^{1/2}$ and $D'^{1/2} \cdot V'^{T}$, respectively. Finally, M and S are obtained by finding a correction matrix A that makes $M' \cdot A$ as orthogonal. The orthographic constraint on M is formulated as

$$\begin{aligned} M' \cdot A \cdot (M' \cdot A)^T &= M' \cdot (A \cdot A^T) \cdot M'^T \\ &= M' \cdot L \cdot M'^T = \begin{bmatrix} 1_{F\times F} & 0_{F\times F} \\ 0_{F\times F} & 1_{F\times F} \end{bmatrix}. \end{aligned} \tag{7}$$

The 3×3 symmetric matrix $L = A \cdot A^T$ with 6 unknown variables is solved first and then $L^{1/2}$ is calculated to obtain

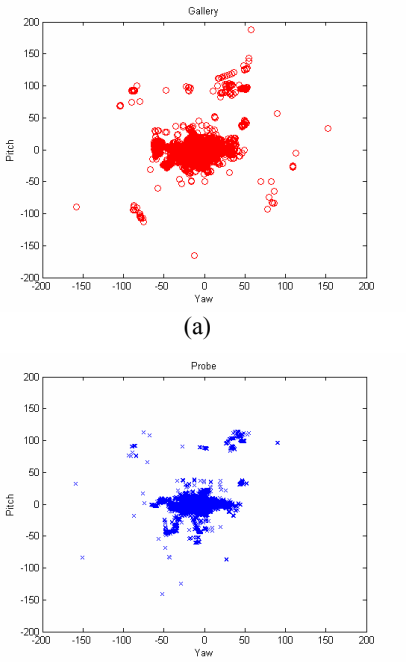


Figure 4: Pose distribution in yaw-pitch space in the (a) gallery and (b) probe data.

A. The Factorization fails i) when the number of frames *F* is less than 2, ii) when singular value decomposition fails or iii) when *L* is not positive definite. Usually, conditions i) and ii) do not occur when processing a video with large number of frames. Most of the failures occur with condition iii). Therefore, the failure condition of Factorization process can be determined by observing the positive definiteness of *L* through Eigenvalue Decomposition.

We estimate the facial pose by iteratively fitting a reconstructed generic 3D model to the facial feature points. The reconstructed 3D shape is first initialized to zero yaw, pitch, and roll, and the iterative gradient descent process is applied to minimize the objective function

$$E = \| P_f - C \cdot R \cdot S \|, \tag{8}$$

where P_f is the 2D facial feature points in the f_{th} frame, C is an orthogonal camera projection matrix, R is the full 3 x 3 rotation matrix, and S is the 3D shape. When the reconstruction process fails, a generic 3D model is used for the pose estimation. The overall process of pose estimation is depicted in Fig. 3. The proposed pose estimation scheme on a synthetic data of 66 frames with known poses varying in the range of [-40°, 40°] with respect to the yaw and pitch shows less than 6° of root mean square error on average. However, this error increases in real face images because of inaccuracies in the feature point detection process. Fig. 4 shows the pose distributions of the probe and gallery data. The total number of images in the probe and gallery set used to generate Fig. 4 are 26,115 and 1,482, respectively. Fig. 4 suggests that there are enough pose variations in both the gallery and probe data mostly in the range of ±60° with respect to the yaw and pitch.

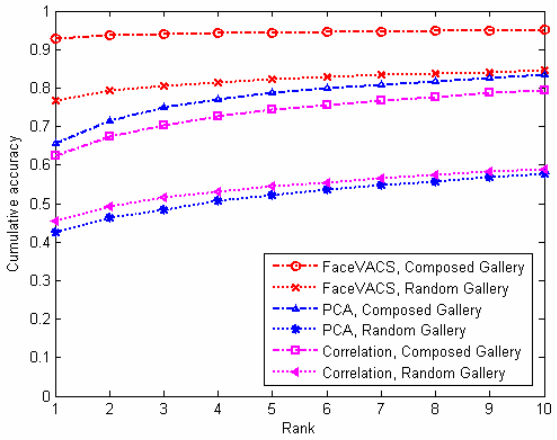


Figure 5: Face recognition performance on two different gallery data; (i) Random Gallery: random selection of pose and motion blur, (ii) Composed Gallery: frames selected based on specific pose and motion blur.

.

5. Motion Blur

Unlike still shot images of the face, motion blur is often present in segmented face images in video. The blurred face images can confound the recognition engine resulting in errors. Therefore, frames with significant motion blur need to be identified and they either need to be selectively enhanced or categorically rejected in the face recognition process. We use Discrete Cosine Transformation (DCT) to estimate and detect motion blur. DCT detects low and high frequency components, and the degree of motion blur can be estimated from the number of high frequency components. We determine the presence of motion blur by observing the DCT coefficients of the top 10% of the high frequency components; frames with motion blur are not considered in the adaptive fusion scheme.

6. Experimental Results

We performed three different experiments to analyze the effect of i) gallery data, ii) probe data and iii) adaptive fusion of multiple matchers on the face recognition performance in video. We first report the experimental results as CMC curves at the frame level. The subject level matching performance is also provided as the overall system performance.

To study the effect of gallery composition, we

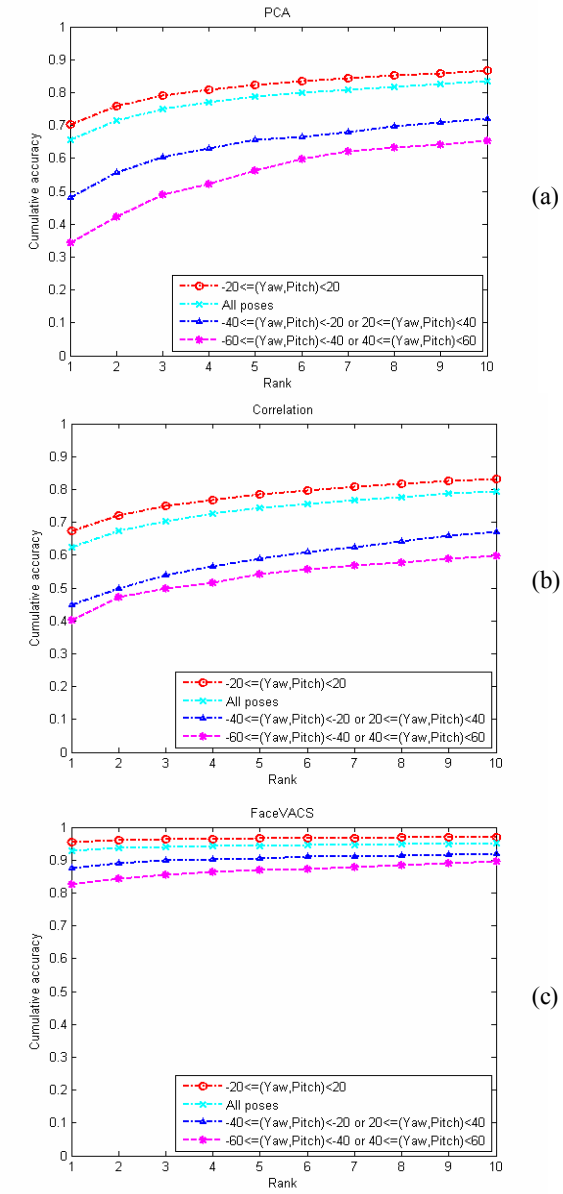


Figure 6: Cumulative matching scores obtained using dynamic information (pose and motion blur) with the three matchers.: (a) PCA, (b) Correlation and (c) FaceVACS

constructed two different gallery datasets. The first gallery set, A, is constructed by selecting 7 frames per subject with pitch and yaw values as {(-40,0), (-20,0), (0,0), (0,20), (0,40), (0,-20), (0,20)}. These frames do not exhibit any motion blur. The second gallery set, B, also has the same number of frames per subject but it is constructed by considering a random selection of yaw and pitch values, and frames in this gallery set may contain motion blur. The effect of gallery dataset on the matching performance is shown in Fig. 5. The gallery database that is systematically composed using pose and motion blur

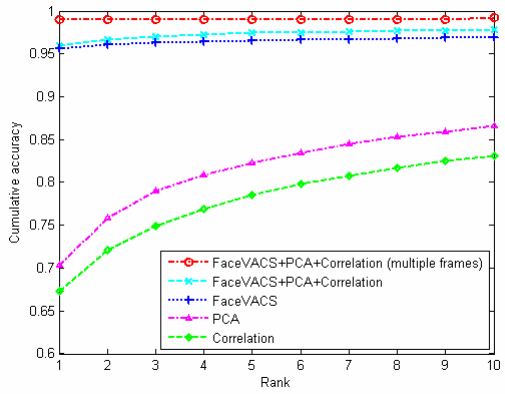


Figure 7: Cumulative matching scores by fusing multiple face matchers and multiple frames in near-frontal pose range (-20< Yaw,Pitch<20).

information (set A) shows significantly better performance across the three matchers. This is because the composed gallery covers the large variations in pose present in the probe data. Removing frames with motion blur also positively affects the performance.

Next, we separate the probe data according to the facial pose in three different ranges: [-20,20], [-40,-20] or [20,40] and [-60,-40] or [40,60] and compute the CMC curves. Fig. 6 indicates that, for all three matchers, the face recognition performance is the best in near frontal-view and decreases when deviating from the frontal view.

Finally, Fig. 7 shows the effects of fusion of multiple matchers and multiple frames using the dynamic information of facial pose and motion blur. We used score-sum with min-max normalization for the matcher level fusion, and max-sum for the frame level fusion. The max-sum decides the identity based on the the maximum score obtained among a number of frames. The best rank-1 accuracy by combining all three matchers is 96%. Frame level fusion (subject level matching accuracy) exhibits an accuracy over 99%. Tables 2 and 3 show examples of matching results of two subjects according to the choice of gallery, probe, and matchers. It is evident from these examples that fusion in the presence of a composed gallery results in the best matching accuracy.

7. Conclusions and Future Work

Face recognition in video has numerous applications, but it also poses a number of challenges. We have shown that the performance of video based face recognition can be improved by fusing multiple matchers and multiple frames in an adaptive manner by utilizing dynamic information pertaining to facial pose and motion blur. This implies that it is crucial to accurately extract dynamic information in the video and use it for face recognition. The systematic use of dynamic information in video is

Table 2: Video example 1: face recognition performance based on gallery, probe and matcher composition.

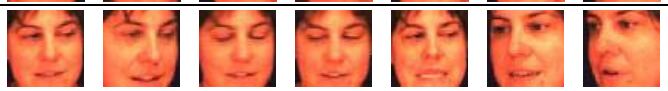
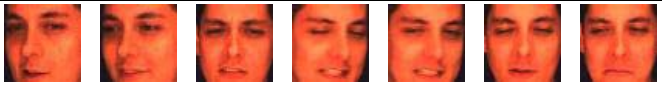
Gallery (random)										
Gallery (composed)		[0,20]	[20,40]	[0,20]	[0,20]	[0,20]	[0,20]	[-40,-20]		
Probe		[-60,-40]	[-60,-40]	[-60,-40]	[-60,-40]	[-60,-40]	[-60,-40]	[-60,-40]	[-40,-20]	[-40,-20]	[-40,-20]
Blurred	yes	yes	no	yes	no	no	no	no	no	no
PCA (random)	1	1	1	1	0	0	0	0	0	0
PCA (composed)	1	1	1	1	1	1	1	1	1	1
Correlation (random)	1	1	1	1	1	0	0	0	0	0
Correlation (composed)	1	1	1	1	1	1	1	1	1	1
FaceVACS (random)	0	0	0	0	1	1	1	1	1	1
FaceVACS (composed)	0	0	0	0	1	1	1	1	1	1
Score-sum (random)	1	1	1	1	1	1	1	1	1	1
Score-sum (composed)	1	1	1	1	1	1	1	1	1	1

Table 3: Video example 2: face recognition performance based on gallery, probe and matcher composition.

Gallery (random)										
Gallery (composed)		[-40,-20]	[-20,0]	[0,20]	[0,20]	[20,40]	[0,20]	[0,20]		
Probe		[0,20]	[0,20]	[20,40]	[20,40]	[20,40]	[20,40]	[20,40]	[20,40]	[20,40]	[20,40]
Blurred	yes	yes	yes	yes	no	no	no	no	no	no
PCA (random)	1	1	0	0	0	0	0	0	0	0
PCA (composed)	0	1	1	0	0	0	0	0	0	0
Correlation (random)	1	0	0	0	0	0	0	0	0	0
Correlation (composed)	0	0	1	1	0	0	0	0	0	0
FaceVACS (random)	1	1	0	0	0	0	0	0	0	0
FaceVACS (composed)	1	1	1	1	0	0	0	0	1	1
Score-sum (random)	1	1	0	0	0	0	0	0	0	0
Score-sum (composed)	1	1	1	1	0	1	0	1	1	1

* Facial pose range shown in square brackets corresponds to the largest value between the yaw and pitch parameters. * 1 and 0 represent a correct and an incorrect match, respectively.

crucial in obtaining a high level of matching accuracy.

The current implementation processes about 2 frames per second, on average. A more efficient implementation of the algorithm and the integration of various modules are underway. As we improve the accuracy of individual modules such as facial pose estimation and 3D shape reconstruction, we also plan to include other dynamic variables (e.g., lighting variations, facial expression, etc.) and additional face matchers (e.g., LDA, LFA, etc.) in this framework.

Acknowledgments. The Face In Action database was

provided by Professor Tsuhan Chen, Carnegie Mellon University. This project was funded in part by the Center for Identification Technology Research (CITeR) at West Virginia University.

References

[1] Stan Z. Li and Anil K. Jain (eds.): Handbook of Face Recognition, Springer, New York, 2005.

[2] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone, Face Recognition Vendor Test 2002: Evaluation Report, Tech. Report NISTIR 6965, NIST, 2003.

[3] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, W. Worek. Preliminary Face Recognition Grand Challenge Results, In Proc. AFGR, pp. 15-24, 2006.

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

[5] Xiujuan Chai, Shiguang Shan, Xilin Chen, and Wen Gao, Local Linear Regression (LLR) for Pose Invariant Face Recognition, In Proc. AFGR, pp. 631-636, 2006.

[6] D. Beymer and T. Poggio, Face Recognition from One Example View, In Proc. ICCV, pp. 500-507, 1995.

[7] V. Blanz and T. Vetter, Face Recognition based on Fitting a 3D Morphable Model, IEEE Trans. PAMI, Vol. 25: 1063-1074, 2003.

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

[9] J Rodney Goh, Lihao Liu, Xiaoming Liu, and Tsuhan Chen, The CMU Face In Action (FIA) Database, In Proc. AMFG, pp. 255-263, 2005.

[10] J. P. Lewis. Fast normalized cross-correlation. Vision Interface, pp. 120-123, 1995.

[11] T. F. Cootes, G.J. Edwords, and C.J. Taylor, Active Appearance Models, Proc. European Conference on Computer Vision, Vol. 2:484-498, 1998.

[12] M. Turk and A. Pentland, Eigenfaces for Face Recognition, Journal of Cognitive Neuroscience, Vol. 3: 71-86, 1991.

[13] P.N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19: 711–720, 1997.

[14] P.S. Penev and J. Attick, Local Feature Analysis: a general statistical theory for object representation, Network: Computation I Neural Systems, Vol. 7: 477-500, 1996.

[15] C. Tomasi and T. Kanade, Shape and motion from image streams under orthography: A factorization method, Int. Journal of Computer Vision, Vol. 9: 137-154, 1992.

[16] S. Romdhani, J. Ho, T. Vetter, and D.J. Kriegman, Face Recognition Using 3-D Models: Pose and Illuminatin, Proc. Of the IEEE, Vol. 94: 1977-1999, 2006.

[17] T. F. Cootes, K. Walker, C. J. Taylor, View-Based Active Appearance Models, Proc. Automatic Face and Gesture Recognition, pp. 227-232, 2000.

[18] J. Kang, I. Cohen, G. Medioni, Continuous Tracking Within and Across Camera Streams, In Proc. Of IEEE CVPR, Vol. 1: I-267-I-272, 2003.

[19] A. F. Bobick and J. W. Davis, The Recognition of Human Movement Using Temporal Templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23: 257-267, 2001.

[20] A. Roy-Chowdhury and Y. Xu, Pose and Illumination Invariant Face Recognition Using Video Sequences, Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems, Springer-Verlag, pp. 9-25, 2006.

[21] S. Zhou, V. Krueger and R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding. Vol. 91: 214-245, 2003.

[22] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden markov models", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1: 340-345, 2003.

[23] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-Based Face Recognition using probabilistic appearance manifolds. In Proc. of Intl. Conf. on Computer Vision and Pattern Recognition, Vol. 1: I-313-I-320, 2003.

[24] G. Aggarwal, A.K. Roy-Chowdhury, R. Chellappa, A System Identification Approach for Video-based Face Recognition, Proc. of the International Conference on Pattern Recognition, Vol. 4: 175-178, 2004.

[25] P. A. Viola, M. J. Jones, Robust Real-Time Face Detection. International Journal of Computer Vision, 57(2): 137-154, 2004.

[26] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott and M. Sharpe, Face Recognition Vendor Test 2006: FRVT 2006 and ICE 2006 Large-Scale Results, Tech. Report NISTIR 7408, NIST, 2007.