Computer Sciences Department

Recognizing Faces from Head Rotation

Guodong Guo
Charles Dyer

Technical Report #1501

May 2004

UNIVERSITY OF WISCONSIN MADISON

Recognizing Faces from Head Rotation

Guodong Guo and Charles R. Dyer
Computer Sciences Department
University of Wisconsin-Madison
Madison, WI 53706

April 1, 2004

Abstract

A new approach for recognizing human faces is presented that uses video sequences of natural, uncontrolled head rotations to capture face motion and dynamic appearance characteristics. Unlike traditional methods for face recognition that utilize one or a few static views, video is used for both training the face recognition system and for recognizing test faces. An algorithm is described that takes an uncalibrated video sequence and extracts the angular rotation of the head in each frame relative to the initial frame. A cropped window of the moving face is also computed, providing a dynamic appearance representation of the face together with the head motion description. Face recognition accuracy using this representation of rotating faces is shown for a small face video database, demonstrating the promise of the method.

1. Introduction

Face recognition has been a research area for over twenty years in computer vision, neuroscience, and psychophysics [2, 18].It is one of the touchstone problems in computer vision research in part due to the wide variety of application areas that depend on this technology, including biometric identification and authentication, and human-computer interaction. Yet, despite many years and many publications, face recognition is not a solved problem. One reason why is that there are many variations on this problem that require different analysis and different algorithms. These variations include assumptions with respect to pose, illumination, facial expression, and external features (e.g., hair styles, glasses, facial hair, and cosmetics).

In order to improve face recognition performance beyond the current state-of-the-art, enriched information to represent each person must be extracted. One promising direction is to use image sequences of moving faces resulting from head movements to extract both motion and dynamic appearance information. Recent psychophysical results show that humans make use of face movement information for recognition [1, 3, 5, 7, 13, 14]. Dynamic information seems to be especially important for humans when spatial image quality is distorted or degraded due to use of photographic negatives, image thresholding, blurring and low resolution [5]. Furthermore, distorting the quality of the motion information, e.g. by slowing down or modifying normal facial motion, negatively impacts recognition performance [7].

In computer vision, there are some publications using video data for face recognition [18]. For example, some work used artificial neural networks [17] [11]. Other approaches used probabilistic techniques to learn the posterior distribution of faces [8, 19] or obtain representative exemplars or cluster centers [6]. In [4], the variations of faces are decoupled or separated from face identities. In [9], a set of views was used to build an "identity surface" with respect to head orientation, and face recognition was done by matching the object trajectory with model trajectories on the identity surfaces. This approach requires that accurate pose be known for each view. In [1] tracked feature point trajectories were used to select a set of key frames, but then the temporal ordering of these chosen frames was maintained so that interpolation between adjacent key-frames could be used for recognition.

Previous video-based face recognition systems do not extract and use head motion information explicitly, although video data has been used as the input either for training or testing. Hence, existing methods rely on still-image based recognition or a slight extension or variation of it.

We present a new approach that uses head rotation information extracted from image sequences from a single camera for both training and testing faces. The key insight in our approach is that head movements are beneficial and should be exploited to obtain a dynamic representation of a face.

The remainder of this paper is organized as follows. Section 2 introduces the main ideas in our approach. Section 3 describes our face video database. Section 4 presents the video processing method used to extract motion and appearance information from an image sequence. Section 5 describes the method for face matching. Experimental re-

sults are given in Section 6.

2. Approach

Consider the following scenario: A face recognition system is being used for person identification in a key-less building entry system. A person walks up and an image is taken of their face. If the system can determine with high confidence the identity of the person based on a single view, then entry is authorized. Otherwise, the person is asked to rotate their head from side to side. Intervals of this continuous sequence of views are then used to resolve the initial low-confidence identity by using these additional views to either increase the confidence of the person's identity sufficiently, or else establish that the person is not a legal user.

This scenario emphasizes several important attributes of our approach that distinguish it from previous work on recognizing faces:

- A face video is represented by a simple frame-indexed description of head rotation plus face appearance
- This representation avoids problems that result when an input image does not closely correspond to any of a set of fixed views

Based on these characteristics, Figure 1 shows the main processing steps for our motion-based approach. The input, whether for learning a new face or recognizing a face, is a video sequence from a single, stationary, uncalibrated camera of a person who is rotating their head. Two modules are then used to extract a motion description and an appearance description, both of which are indexed by frame number in the sequence. The database stores for each video of a person, a sequence of (head orientation, face template) pairs, one for each frame in the input sequence. A face to be recognized is processed identically, and the recognition phase then compares the similarity of face templates at corresponding head orientations.

3. A Face Video Database

To demonstrate this new framework, we have built a small face video database. In deciding what type of head motion to capture, there are several important objectives: (1) the motion should be natural for people, (2) the motion should contain enough information to improve face recognition accuracy, (3) the motion should not be too complex to extract motion features, (4) the system should not require customized cameras or a restrictive studio environment, and (5) the system should not require calibrated cameras or accurate positioning of the head. Furthermore, in order to include analysis of fine motion details, we use close-up views and high-resolution images of the individuals' faces so that the acuity of facial features is high.

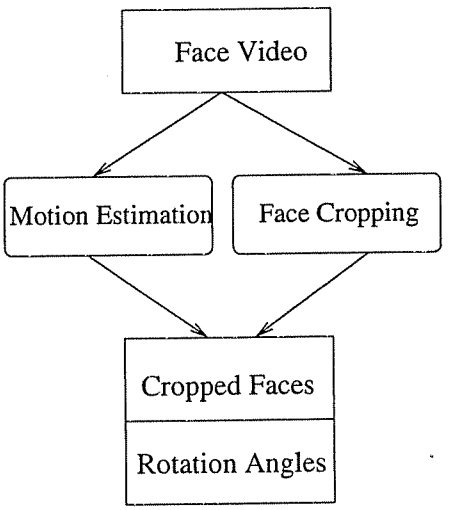

Figure 1: An overview of our approach to the input face video.

Based on the above considerations and psychophysical studies, horizontal head rotation is a good choice for capturing key face motion properties. We ask each subject to rotate his or her head from an approximately frontal view to an approximately profile view (i.e., approximately a 90 degree head rotation). A single, stationary, uncalibrated camera is used to capture video of a subject. Currently, there are 18 individuals captured, each with 3 videos. The number of frames per video varies, ranging from 133 to 199, and each image is size 720×480. In total there are 10,182 images in our face video database. Some example images from one of our face videos are shown in Figure 2. Finally, note that while there is an existing face video database called XM2VTS [12], it contains complex motions (both left to right and up-down) and some motion blur, so it was not appropriate for our work.

4. Face Video Processing

4.1. Background Removal

To simplify the preprocessing phase, our studio has a blue screen that subjects sit in front of, so that simple segmentation techniques can be used to separate the person from the background. To deal with variable illumination, we first compute the normalized blue component $b = B/(R+G+B)$ at every pixel. Next, using a small window that is known to be part of the background, e.g., in the upper-right corner of the image, the mean, μ, and standard deviation, σ, of the normalized blue component values in the window are computed. The background is labeled as those pixels, i, where $b_i > \mu - 5\sigma$. An example of the results of this segmentation step, which is applied independently to every frame in

Figure 2: Selected images from a rotating head sequence in our face video database.

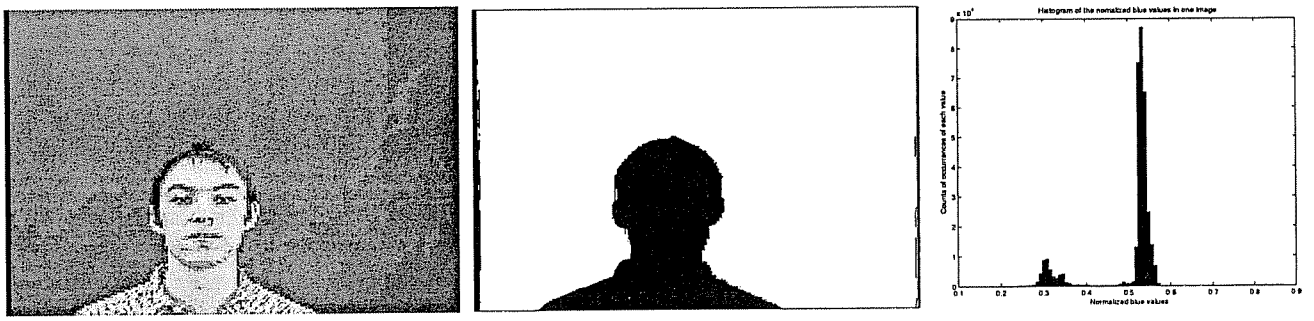

Figure 3: Result of background segmentation. One frame of a video is shown on the left. The histogram shows the distribution of the normalized blue components in the input image. The middle image shows the computed background mask as white.

an input video, is shown in Figure 3.

4.2. Face Tracking

Once the background pixels have been removed, it is necessary to track the motion of the head. Since we do not make any special assumptions about the motion at this stage, we use the off-the-shelf KLT tracker [15] for this purpose. Because of self-occlusion as a head rotates, we run the KLT tracker on consecutive intervals of n frames, where $n = 50$ in our experiments. Background pixels are not considered as feature points for tracking, which means most of the tracked points are part of the head, though some outliers occur from points tracked on the person's neck and shoulders.

4.3. Head Motion Estimation

Using the tracked feature points across consecutive frames, we can now estimate the head motion. We use the factorization method [16] to describe the motion of facial feature points. That is, given point correspondences from the tracker, we build a measurement matrix and use singular value decomposition (SVD) to solve for the camera motion. Camera motion is represented by the image plane coordinate directions, i and j. In our case, the camera is fixed and the object is rotating about an approximately vertical axis. The angle between two image planes (i_m, j_m) and (i_n, j_n) can be computed by a dot product between the two normal vectors N_m and N_n, i.e., $\theta_{mn} = \arccos(N_m \cdot N_n)$, where $N_k = i_k \times j_k$ for $k = m, n$. If we take the first image as the reference frame, the angle between any other frame and the reference can be estimated. Because we only need to compute head rotation, only the motion matrix component of the factorization method is used; we do not use the shape matrix that is part of the traditional factorization method.

4.4. Outlier Removal

The factorization method is based on point correspondences and therefore its results depend on the quality and consistency of the feature points used. Because of occlusion, illumination variation, and non-moving parts, we can not expect the KLT tracker to use only feature points that are part of the moving head, and some correspondences may not always be correct. In Section 4.1 background pixels were removed to partially deal with this issue. However, feature points detected on the person's neck or shoulders will not undergo the same motion as the head, so these outlier points need to be detected and removed in order to make the rotation estimation more accurate.

As an initial prototype, we implemented three filters to remove outliers:

- Remove feature point correspondence errors. Because the 3D motion is known to be continuous throughout the sequence (another benefit of using a video instead of several widely-separated views), the 2D projection of any feature point that is tracked should change smoothly through the sequence. If two corresponding points in two consecutive frames have a large change in their x or y coordinates, a correspondence error is detected and that feature point is removed.

- Remove stationary feature points. Some points belonging to other parts of the person's body such as their neck and shoulders will be detected and tracked by the KLT tracker. One simple method for detecting these points is to compute the maximum horizontal range of motion of each point over the entire sequence, i.e., $maxX$ and $minX$ (remember the head rotates about a vertical axis), and if $|maxX - minX|$ is less than a threshold, the point is eliminated.

- Remove points that become occluded during the sequence. In our motion sequences the head rotates to the right, so feature points on the right side of the face become occluded as the head moves from its initial frontal position. Consequently, we detect and remove about 5 percent of the feature points that are on the rightmost side of the face.

The above three outlier removal filters are critical for obtaining reliable motion information. Figure 4 demonstrates this point by showing the rotation angles computed with and without outliers included. Some image frames from the sequence used for this example are shown in Figure 2. The results in Figure 7 were also obtained after removing outliers.

4.5. Face Region Cropping

As a final step, each image is tightly cropped so that the face predominates, simplifying subsequent appearance-based recognition. Face cropping and face segmentation are themselves difficult problems and methods based on facial features such as eye and mouth positions have been used previously for this purpose on frontal face images [10]. These approaches are not appropriate in our case because of the wide range of poses that must be accommodated. Instead, we have used the following simple procedure for cropping each frame in a sequence:

1. Compute the top-of-head position in the first frame using the background mask produced by blue-screen detection.

2. Compute the head width in the first frame from the background mask.

3. Determine the head height. An aspect ratio α is set by trying multiple values in the range $[0.9, 1.1]$ and selecting the best one.

4. Adjust the window position at successive frames, using the same size window, but translated to best fit the background mask.

5. Resize the cropped windows to a fixed size.

5. Face Matching

The result of the steps for processing a video sequence described in the last section is, for each frame in each video, (1) the estimated head motion, given as an angle of rotation relative to the first frame's orientation, and (2) a cropped window of the face. Before using this representation of motion and appearance for recognition, we consider two questions: Is the motion estimation correct, and is the face region cropped correctly?

One way to evaluate the results of the video processing stage is to compare two face videos at a selected rotation angle θ and verify whether the rotation angles are estimated correctly and whether the face templates are cropped correctly at that pose.

We define similarity measure, D, between two videos, V_{test} and V_{train}, with respect to a given angle θ, using the RGB color images from the sequences and 1-norm distance, as

$$\begin{aligned} & D(V_{test}, V_{train}, \theta) = D(T_i(V_{test}, \theta_i), T_j(V_{train}, \theta_j)) \\ = & \parallel T_i(V_{test}, \theta_i, R) - T_j(V_{train}, \theta_j, R) \parallel_1 \\ + & \parallel T_i(V_{test}, \theta_i, G) - T_j(V_{train}, \theta_j, G) \parallel_1 \\ + & \parallel T_i(V_{test}, \theta_i, B) - T_j(V_{train}, \theta_j, B) \parallel_1 \end{aligned}$$

where θ_i and θ_j are the closest angles to the given angle θ in the test video and the training video, respectively, and

$$\begin{aligned} i &= \arg\min_k \parallel \theta_k - \theta \parallel, \quad \theta_k \in \Theta_{test} \\ j &= \arg\min_m \parallel \theta_m - \theta_i \parallel, \quad \theta_m \in \Theta_{train} \end{aligned}$$

where Θ_{test} and Θ_{train} are the sets of rotation angles estimated for each frame in the test and training videos, V_{test} and V_{train}, respectively. R, G, B are the red, green, and blue pixel values in the cropped face templates. See Fig. 5 for a graphical interpretation of how a test template T_i is related to a corresponding training template T_j via the rotation angle indices, θ_i and θ_j. Note also that the training and test data may have different lengths.

6. Experimental Results

In this section we describe two experiments that demonstrate the potential of our approach. First, we verify the robustness of the head motion estimation process by using a controlled mannequin head sequence where the ground truth rotation angles are known. The second experiment shows results of motion estimation and face template extraction in real face videos.

6.1. Motion Estimation for a Mannequin Head

The question of how to accurately compute head motion from a face video is of critical importance to our method.

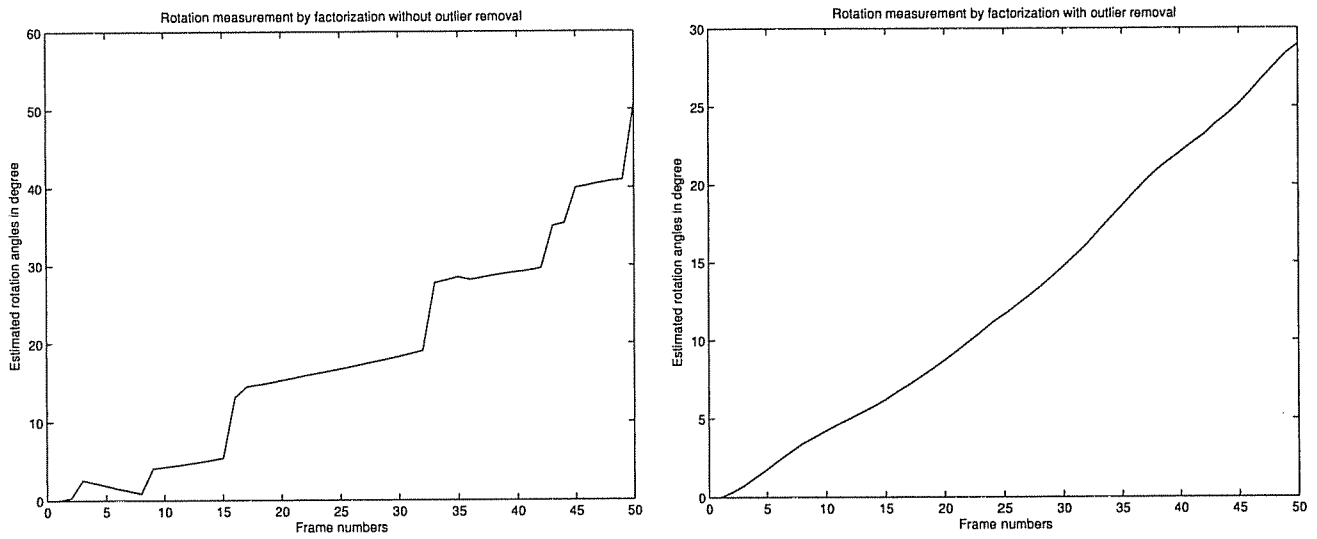

Figure 4: Rotation angles computed for the first 50 frames of a face video without outlier removal (left), and with outlier removal (right).

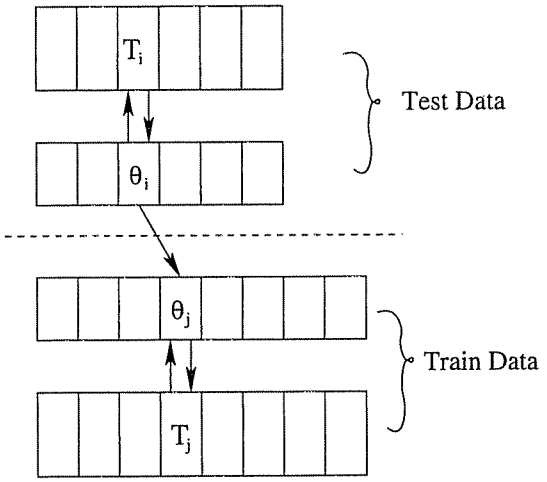

Figure 5: A depiction of how corresponding face templates in two different videos are indexed using associated rotation angles.

As described in Section 4.3, we use the factorization method to compute the angle between a given frame and the first frame. To verify the robustness and accuracy of this SVD-based technique for head motion estimation, we used a pan-tilt unit as a turntable with a mannequin head mounted on it. The mannequin head was rotated under computer control by 1 degree increments and an image taken at each position. Figure 6 shows images of the mannequin head taken at several rotation angles.

The estimated head motion, i.e., the rotation angles relative to the first frame are shown in Figure 7. The fact that the slope of the curve is nearly constant is consistent with the known, uniform rotation of the mannequin head. Quantitatively, the maximum error of the computed angle was 2.3 degrees at frame 57.

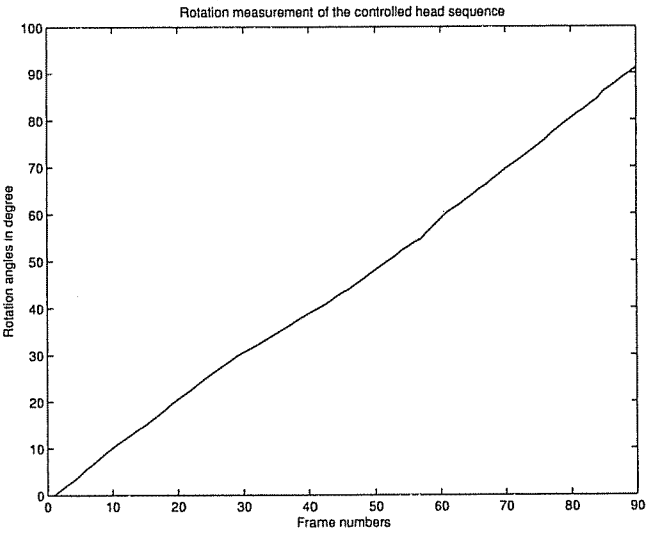

Figure 7: Rotation angles computed for the mannequin-head motion sequence.

6.2. Motion Estimation, Face Cropping, and Face Recognition for Real Videos

In this section we describe experiments using the face video database described in Section 3, which contains 18 individ-

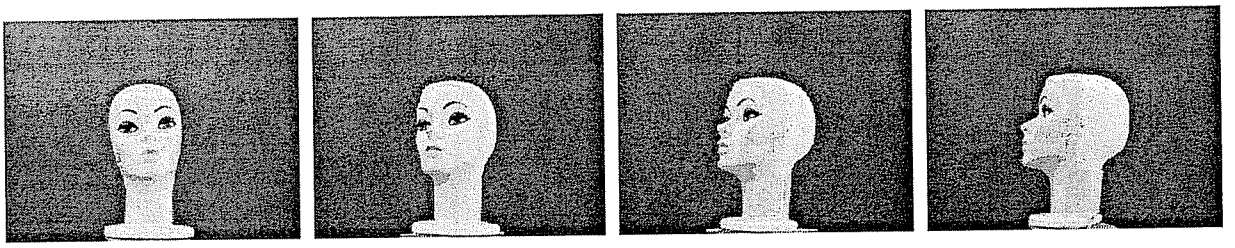

Figure 6: A mannequin head at orientations 0, 30, 60, and 90 degrees.

uals, with three image sequences for each person, for a total of 54 videos. Each frame is size 720 × 480, and the cropped face templates are 128 × 128. Some of the results of video processing to compute the cropped faces and rotation angles are shown in Figure 8. The computed head rotation angles are also given for visual verification.

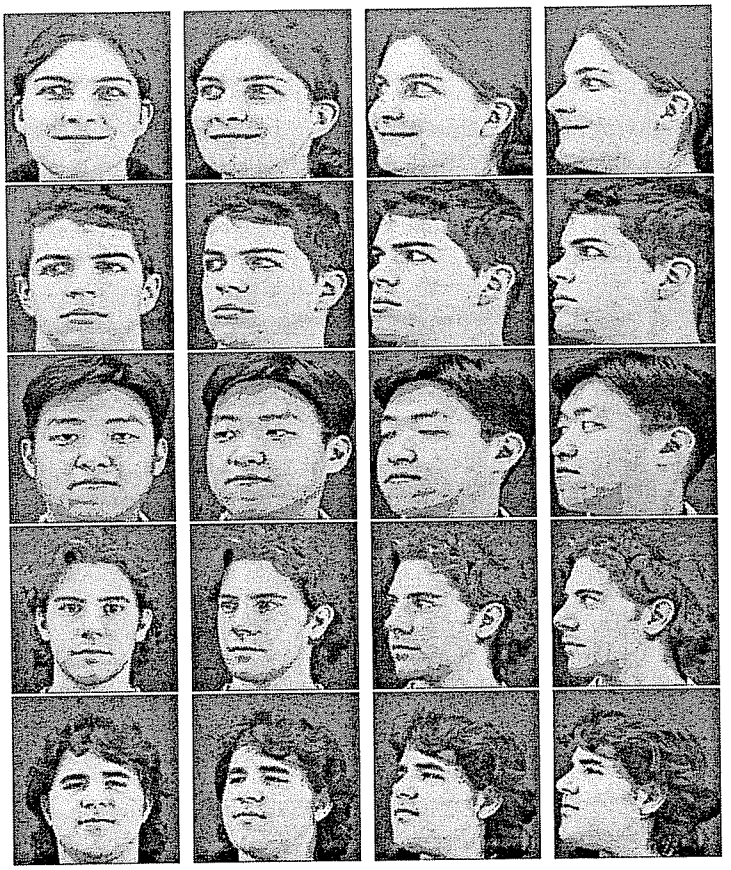

Figure 8: Some sample results of the face cropping procedure. The columns correspond to frames whose associated rotation angles were computed to be approximately 0, 20, 40, and 60 degrees.

As an initial evaluation of recognition performance, we randomly selected one video for each person as the training example, so the training set contained 18 image sequences. The other two videos for each person were used in the test set, which therefore was of size 36. We selected test images from each of the 36 test videos corresponding to rotation angles from 0 to 60 degrees at intervals of 5 degrees. So, a total of $2 \times 18 \times 13 = 468$ cropped face images were used as the test set. (Note: Because it is difficult for many people to easily rotate their head more than 70 degrees, we only used test images up to 60 degrees.) For a given test video the frame corresponding to rotation angle θ was determined by finding the closest angle, θ_i, in the test video such that $i = \arg\min_k \parallel \theta_k - \theta \parallel, \forall \theta_k \in \Theta_{test}$, where Θ_{test} is the set of angles computed for all frames in the test video. Similarly, the closest angle θ_j to θ in each training video was found, and its corresponding cropped face template T_j selected. Test template T_i and training template T_j were compared using the similarity measure defined in Section 5. The training video whose template is most similar to the test template T_i was selected as the identified person.

Using this procedure with the 468 test images and the 18 training videos resulted in 100% correct classification. This suggests that (1) the head rotation angles were estimated well, and (2) the face templates were sufficiently well-cropped in each frame. Furthermore, this indirectly indicates that the faces at various orientations other than frontal are also useful for face recognition.

Finally, notice that it is possible to use this approach when the frontal views of two people can not be discriminated. So, one possible extension is to compare neighboring face templates $T_{i'}$ and $T_{j'}$, indexed by $\theta_{i'}$ and $\theta_{j'}$, when the similarity between T_i and T_j is less than a confidence level. Furthermore, checking multiple values of i' and j', say at a specified interval, could be done to also improve the confidence of the person identified.

6.3. Discussion

In this paper two assumptions were made that are important to comment upon. One is the orthographic projection assumption used by the factorization method. The other is the frontal view assumption for the first frame in each video. Because both assumptions are not true for real data, they both introduce errors in the computation of head orientation. In particular, because each person controls their own

initial head position, the first frame will rarely correspond to a true 0 degree orientation with respect to the camera's optical axis. Despite these errors, our experimental results show that the estimates of rotation angle need not be highly accurate in order to obtain high recognition accuracy. In fact, for faces at least, recognition performance is usually not very sensitive to small pose variations, e.g., less than 5-10 degrees. This has been observed in earlier results for still-image based face recognition.

7. Concluding Remarks

This paper presented a new approach for recognizing human faces from videos of natural head rotations. Our approach is computationally efficient because it is based on fast, standard techniques for extracting motion and appearance information, combined with simple indexing and correlation-based face matching. The representation is also potentially very space efficient because the cropped face templates in a video are highly correlated and therefore can be highly compressed. Finally, because the method extends naturally to allow for iteratively increasing the confidence of recognition by testing additional frames, it promises to scale up well to large databases.

References

[1] H. H. Bülthoff, C. Wallraven, and A. Graf. View-based dynamic object recognition based on human perception. *Proc. 16th Int. Conf. Pattern Recognition, Vol. 3*, pages 768–776, 2002.

[2] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proc. IEEE*, 83:705–741, May 1995.

[3] F. Christie and V. Bruce. The role of dynamic information in the recognition of unfamiliar faces. *Memory and Cognition*, 26:780–790, 1998.

[4] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Improving identification performance by integrating evidence from sequences. *Proc. Computer Vision and Pattern Recognition Conf.*, pages 486–491, 1999.

[5] B. Knight and A. Johnston. The role of movement in face recognition. *Visual Cognition*, 4(3):265–273, 1997.

[6] V. Kruger and S. Zhou. Exemplar-based face recognition from video. *Proc. 7th European Conf. Computer Vision, Vol. IV*, pages 732–746, 2002.

[7] K. Lander and V. Bruce. Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology*, 12(4):259–272, 2001.

[8] B. Li and R. Chellappa. Simultaneous tracking and verification via sequential posterior estimation. *Proc. Computer Vision and Pattern Recognition Conf., Vol. 2*, pages 110–117, 2000.

[9] Y. Li, S. Gong, and H. Liddell. Constructing facial identity surfaces in a nonlinear discriminating space. *Proc. Computer Vision and Pattern Recognition Conf., Vol. 2*, pages 258–263, 2001.

[10] A. M. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(6):748–763, 2002.

[11] A. Massad, B. Mertsching, and S. Schmalz. Combining multiple views and temporal associations for 3-D object recognition. *Proc. 5th European Conf. Computer Vision, Vol. II*, pages 699–715, 1998.

[12] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. *Proc. 2nd Int. Conf. Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[13] A. J. O'Toole, D. A. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science*, 6(6):261–266, 2002.

[14] G. E. Pike, R. I. Kemp, N. A. Towell, and K. C. Phillips. Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition*, 4(4):409–437, 1997.

[15] J. Shi and C. Tomasi. Good features to track. *Proc. Computer Vision and Pattern Recognition Conf.*, pages 593–600, 1994.

[16] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. Computer Vision*, 9(2):137–154, 1992.

[17] H. Wechsler, V. Kakkad, J. Huang, S. Gutta, and V. Chen. Automatic video-based person authentication using the RBF network. *Proc. 1st Int. Conf. Audio and Video-Based Biometric Person Authentication*, 1997.

[18] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips. Face recognition: A literature survey. Technical Report CAR-TR-948, Center for Automation Research, University of Maryland, 2002.

[19] S. Zhou and R. Chellappa. Probabilistic human recognition from video. *Proc. 7th European Conf. Computer Vision, Vol. III*, pages 681–697, 2002.