Frame Synchronization and Multi-Level Subspace Analysis for Video Based Face Recognition

Xiaoou Tang and Zhifeng Li

Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
{xtang, zli0}@ie.cuhk.edu.hk

Abstract

In this paper, we develop a new video-to-video face recognition algorithm. The major advantage of the video based method is that more information is available in a video sequence than in a single image. In order to take advantage of the large amount of information in the video sequence and at the same time overcome the processing speed and data size problems we develop several new techniques including temporal and spatial frame synchronization and multi-level subspace analysis for video cube processing. The method preserves all the spatial-temporal information contained in a video sequence. Near perfect classification results are obtained on the XM2VTS face video database.

1. Introduction

Automatic face recognition is a challenging task in pattern recognition research. In recent years, a number of techniques have been proposed including local feature analysis methods such as the Active Appearance Model (AAM) [14] and the elastic graph matching (EGM) method [5] and the appearance-based subspace methods such as the eigenface method [4], the LDA method [1][6], and the Bayesian algorithm [2]. Many of these methods and their combinations have shown promising recognition performance in the FERET test [3].

However, all of these methods focus exclusively on image-based face recognition that uses a still image as input data. One problem with the image-based face recognition is that it is possible to use a pre-recorded face photo to confuse a camera to take it as a live subject. The second problem is that the image-based recognition accuracy is still too low in some practical applications comparing to other high accuracy biometric technologies. To alleviate these problems, video based face recognition has been proposed recently [8][9][10][11]. One of the major advantages of video-based face recognition is to prevent the fraudulent system penetration by pre-recorded facial images. The great difficulty to forge a video sequence (possible but very difficult) in front of a live video camera may ensure that the biometric data come from the user at the time of authentication. Another key advantage of the video based method is that more information is available in a video sequence than in a single image. If the additional information can be properly extracted, we can further increase the recognition accuracy.

However, contrary to the large number of image-based face recognition techniques, the research on video-to-video face recognition has been limited. Most research on face recognition in video has mainly been focusing on face detection and tracking in video. Once a face is located in a video frame, the conventional image based face recognition technique will be used for a single frame recognition. For recognition directly using video data, Satoh [8] matches two video sequences by selecting the pair of frames that are closest across the two videos, which is inherently still image-to-image matching. Methods in [9][10] use video sequence to train a statistical model face for matching. Even though the trained model is more stable and robust than a model trained from a single image, the overall information contained in the model is still similar to a single image given the same feature dimension. This is similar to image-to-image matching with increased training data size. The mutual subspace method in [8][11] uses the video frames for each person separately to compute many individual eigenspaces. Since it cannot capture discriminant information across different people, the recognition accuracy is lower than other methods.

In this paper, we propose a new video-to-video face recognition algorithm that takes full advantage of the complete spatial temporal information contained in a video sequence. Although more information is available in a video sequence than a single image, and thus may help to increase the recognition accuracy, this advantage comes at a cost. More data means more information, at the same time, means higher processing complexity. In order to extract discriminant information efficiently from video sequence for face recognition, we have to

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

overcome several key hurdles of processing speed and large data size.

First we develop a video frame temporal synchronization method. The idea is to align frames of similar images across the two video sequences so that they can be better matched. Given the large amount of data in video, we cannot afford to use a complicated algorithm for this purpose. We propose a very simple and effective algorithm taking advantage of the audio signal in video. We use the waveform of the audio signal to allocate desired frames in each video. After the temporal synchronization, we conduct spatial synchronization by aligning key fiducial points on each image using Gabor wavelet feature [5]. Alignment of the fiducial points is critical for subspace methods to take advantage of the shape correlation across different face images. Finally, for fast matching of the large spatial and temporal synchronized video sequence, we develop a multi-level subspace analysis algorithm. Experiments on the largest standard video face database, the XM2VTS database [7], show near perfect recognition accuracy.

2. Video Frame Synchronization

In video based recognition, for the video to provide more information, individual frames in a video have to be different from each other. Since if all the frames are similar to each other, the information contained in the video sequence will be basically the same as a single image. However, for videos of varying frame contents, a simple matching of the two video sequence frame by frame will not help much, since we may be matching a frame in one video with a frame of different expression in another video. This may even deteriorate the face recognition performance.

The key for the performance improvement is that the images in the sequence has to be in the same order for each individual, so that neutral face matches with neutral face and smile face matches with smile face. Therefore, if we want to use video sequence for face recognition, it is important to synchronize similar video frames in different video sequence. We call this "temporal synchronization" since we will re-order the original temporal video sequence according to different content in each frame. To accomplish this we can use regular image-based expression recognition techniques to match similar expression in different video. However, the computation is too costly for the large amount of video data. The expression recognition accuracy is also not very high. Here we propose a new approach using information in the audio signal in the video.

For example, the video data in the XM2VTS database (the largest publicly available face video database) contains video sequences for 295 people. For each person, several video sequences of 20 seconds each are taken over four different sessions. In each session, a person is asked to recite two sentences "0, 1, 2, …, 9" and "5, 0, 6, 9, 2, 8, 1, 3, 7, 4" when recording the video sequences. We can use these speech signals to locate frames with distinctive expressions. An example is shown in Fig. 1, where we locate the maximum point of each word and select the corresponding video frames. We can see different expressions when one read different word. Of course more sophisticated speech recognition technique can also be used to improve the result with added computational cost. We found our simple approach already very effective and efficient and is good enough for our recognition purpose. The audio-guided method helps us to synchronize video sequence and select a number of distinctive frames for face recognition. In addition, the method can be easily extended to include more speech information. For example, speaker verification based on the user's voice and verbal information verification based on the message content can also be integrated with the video sequence to achieve better performance.

After the temporal synchronization, the next step is to align key fiducial points on each image since when people are talking, their face will move and change. We call this step spatial synchronization. Alignment of the fiducial points is critical for subspace methods to take advantage of the shape correlation across different human faces. We use the Gabor wavelet feature [5] to allocate key fiducial points for the spatial synchronization.

3. Multi-level Subspace Analysis

After the spatial and temporal synchronization, we finally have an aligned 3D face data cube for each person. There are a number of ways that we can conduct the video sequence matching. As discussed earlier, using traditional methods such as nearest image or mutual subspace methods cannot utilize all the discriminant information in the video data. A straightforward approach is to treat the whole data cube as a single large feature vector and conduct regular subspace analysis to extract features. Although this feature level fusion approach utilized all the data in video, there are several problems with this approach. First, the data size will be extremely large. In our experiments, we use 21 images of size 41x27 for each video sequence, thus the feature dimension is 23247. Direct subspace analysis on such a large vector is too costly. Second, a more serious problem is the over fitting problem because of the small sample size versus large feature dimension for discriminant subspace analysis algorithms.

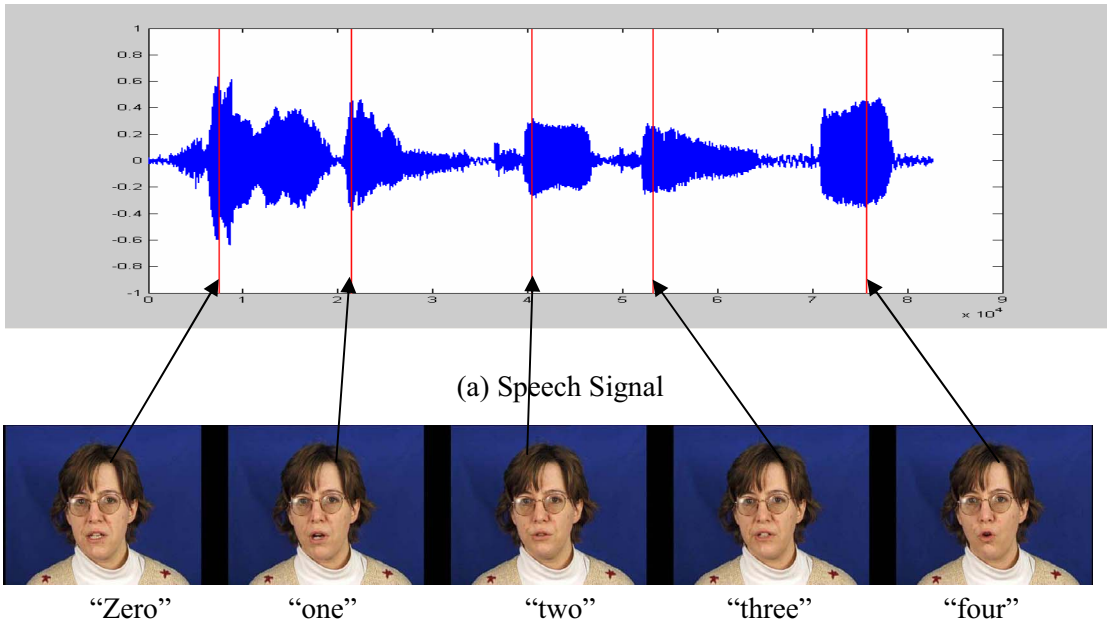

Figure. 1. Example video sequence and corresponding speech signal.

To overcome these problems, we develop a multi-level subspace analysis algorithm. We first break the video cube into slices, with features from each frame as a slice. Then we perform unified subspace analysis [12] on each feature slice. The extracted discriminant features from each slice are then combined to form a new feature vector. We then apply PCA to the new feature vector to remove redundant information among the feature slices to extract the final feature vector. The detail algorithm is as follows.

In the first level subspace analysis, for each feature slice:

1. Project each feature slice to its PCA subspace computed from the training set of the slice and adjust the PCA dimension to reduce most noise.
2. Compute the intrapersonal subspace using the within-class scatter matrix in the reduced PCA subspace and adjust the dimension of intrapersonal subspace to reduce the intrapersonal variation.
3. For the L individuals in the gallery, compute their training data class centers. Project all the class centers onto the intrapersonal subspace, and then normalize the projections by intrapersonal eigenvalues to compute the whitened feature vectors.
4. Apply PCA on the whitened feature vector centers to compute the final discriminant feature vector.

In the second level of subspace analysis,

1. Combine the extracted discriminant feature vectors from each slice into a new feature vector.
2. Apply PCA on the new feature vector to remove redundant information in multiple frames. The features with large eigenvalues are selected to form the final feature vector for recognition.

In the second level subspace analysis we only use PCA instead of unified subspace analysis. This is because the intrapersonal variation has already been reduced in the first level whitening step and discriminant features have been extracted in step 4 of the first level. Repeating them will not add any new information. However, there is a significant amount of overlap information between different slices since the frames are still quite similar with each other even with expression changes. PCA is needed to reduce the redundant information.

We can show that the multiple level subspace analysis does not loss much information compared to the original subspace analysis. This is similar to the multilevel dominant eigenvector estimation algorithm in [13]. Here we gave a more detailed proof. Since the whitening step removes intra-personal variations which contain only unwanted information, we do not need to consider them when analyzing information lose in the algorithm. So we only need to focus on the two PCA steps. To compute PCA, we first form an n by m sample matrix,

$$A = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_m(1) \\ x_1(2) & x_2(2) & \dots & x_m(2) \\ \dots & \dots & \dots & \dots \\ x_1(n) & x_2(n) & \dots & x_m(n) \end{bmatrix}, \qquad (1)$$

where x_i is a video cube feature vector, n is the vector length, and m is the number of training samples. By breaking the long feature vector into g = n/k groups of small feature slices of length k,

$$A=\begin{bmatrix} B_1\left\{\begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_m(1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(k) & x_2(k) & \cdots & x_m(k) \end{bmatrix}\right\} \\ B_2\left\{\begin{bmatrix} x_1(k+1) & x_2(k+1) & \cdots & x_m(k+1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(2k) & x_2(2k) & \cdots & x_m(2k) \end{bmatrix}\right\} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ B_g\left\{\begin{bmatrix} x_1((g-1)k+1) & x_2((g-1)k+1) & \cdots & x_m((g-1)k+1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(n) & x_2(n) & \cdots & x_m(n) \end{bmatrix}\right\} \end{bmatrix} \quad , (2)$$

we can perform PCA on each of the g group short feature vector set B_i. Then a new feature vector is formed by the first few selected eigenfeatures of each group. The final eigenvectors are computed by applying PCA to this new feature vector. To prove that the eigenvalues computed this way are a close approximation of the standard one step PCA, we study the two-group case here. The feature vector matrix and its covariance matrix are

$$A=\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \tag{3}$$

$$W = AA^T = \begin{bmatrix} B_1 B_1^T & B_1 B_2^T \\ B_2 B_1^T & B_2 B_2^T \end{bmatrix} = \begin{bmatrix} W_1 & W_{12} \\ W_{21} & W_2 \end{bmatrix}. \tag{4}$$

Let the eigenvector matrices of the covariance matrices W_1 and W_2 be T_1 and T_2 respectively, then

$$T_1^T W_1 T_1 = \Lambda_1, \tag{5}$$

$$T_2^T W_2 T_2 = \Lambda_2, \tag{6}$$

Where Λ_1 and Λ_2 are the diagonal eigenvalue matrices. The effective rotation matrix for the first-step group PCA is

$$T=\begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}. \tag{7}$$

T is also an orthogonal matrix, since

$$T^T T = \begin{bmatrix} T_1^T T_1 & 0 \\ 0 & T_2^T T_2 \end{bmatrix} = I. \tag{8}$$

So, after the first-step group PCA, the covariance matrix of the rotated feature vector,

$$W_r = T^T W T = \begin{bmatrix} \Lambda_1 & T_1^T W_{12} T_2 \\ T_2^T W_{21} T_1 & \Lambda_2 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \Lambda_{1b} & 0 \\ 0 & \Lambda_{1s} \end{bmatrix} & \begin{bmatrix} C_{bb} & C_{bs} \\ C_{sb} & C_{ss} \end{bmatrix}^T \\ \begin{bmatrix} C_{bb} & C_{bs} \\ C_{sb} & C_{ss} \end{bmatrix} & \begin{bmatrix} \Lambda_{2b} & 0 \\ 0 & \Lambda_{2s} \end{bmatrix} \end{bmatrix} \quad , (9)$$

is a similar matrix of the original feature vector covariance matrix W, because of the orthogonality of the rotation matrix T. Since similar matrices have the same eigenvalues, we can use the right most term of Eq. (9) to discuss the impact on W by keeping only the first few dominant eigenvalues in each group. In Eq. (9), Λ_{nb} and Λ_{ns} represent the larger dominant eigenvalue section and the smaller negligible eigenvalue section of the eigenvalue matrix Λ_n respectively, for $n=1$ or 2. C_{xx}, where $x=b$ or s, represents the cross-covariance matrix of the two groups of rotated features. By keeping only the dominant eigenvalues in the second level PCA, the new feature vector covariance matrix becomes

$$W_d = \begin{bmatrix} \Lambda_{1b} & C_{bb}^T \\ C_{bb} & \Lambda_{2b} \end{bmatrix}. \tag{10}$$

The terms removed from W_r are Λ_{1s}, Λ_{2s}, C_{ss}, C_{bs}, and C_{sb}. Since most energy is contained in the dominant eigenvalues, the loss of information due to Λ_{1s} and Λ_{2s} should be very small. The energy contained in the cross-covariance matrix of the two small energy feature vectors, C_{ss}, should therefore be even smaller.

We can also show that C_{bs} and C_{sb} cannot be large either. If the two group features B_1 and B_2 are fairly uncorrelated with each other, then all the cross-covariance C_{xx} matrices in Eq. (9) will be very small. On the other hand, if the two group features are strongly correlated with each other, the dominant eigenfeatures of the two groups will be very similar. Therefore the cross-covariance matrix C_{bs} of group-two large features with group-one small features will be similar to the cross-covariance matrix of the group-one large features with group-one small features, which is zero due to the decorrelation property of PCA.

When the two group features B_1 and B_2 partially correlated, the correlated part should be mostly signal, since noise parts of the variable B_1 and B_2 rarely correlate with each other. The basic property of PCA is to preserve all signal energy in the first few large

eigenvalues. Therefore, most signal energy in B_2, and especially most of the B_2 signal energy that is correlated with B_1, will be preserved in the large eigenvalue section of B_2 covariance matrix. The energy that is discarded in the small eigenvalue section of B_2 will contain little if any energy that is correlated with B_1. Therefore, C_{bs} and C_{sb} should be very small, and we will not lose much information by removing them from the covariance matrix W_r.

Now that we have shown that the covariance matrix W_d is a close approximation of W_r, and W_r is a similar matrix of W, we can say that the eigenvalues from W_d of the multi-level subspace method, are indeed a close approximation of the eigenvalues computed from W of the standard PCA method.

4. Experiments

In this section, we conduct experiments on the XM2VTS face video database [7]. We select 294*4 video sequences of 294 distinct persons from the four different sessions. For the training data, we select the 294*3 video sequences of the first three sessions. The gallery set is composed of the 294 video sequences of the first session. The probe set is composed of the 294 video sequences of the fourth session. The persons in the video are asked to read two sequence of numbers, "0 1 2 3 4 5 6 7 8 9" and "5 0 6 9 2 8 1 3 7 4".

From each video, 21 frames are selected by means of two strategies respectively: Audio-Video Temporal Synchronization and random selection without the audio information. So there are two different sets of face image sequences labeled as A-V Synchronization data and A-V non-synchronization data respectively. For the A-V synchronization data, each frame corresponds to the waveform peak of a digit. An additional frame is located at the midpoint of the end of the first sentence and the start of the second sentence.

We first look at the recognition results of appearance based methods using image gray scale values directly as features. The results for both still image and video sequence are summarized in Table 1. The still image is either selected from the first frame of the video sequence (A-V Synchronization case), or is selected randomly from the video sequence (A-V Non-Synchronization case). We can see that the performance of using still image directly by Euclidean distance classification is very poor (61%). This baseline result actually reflects the difficulty of the database. As we know that for face recognition experiments, if the probe image and the gallery image are from different sessions, the result is usually poor. This is the case for our experiments. Significant improvement is achieved by video data using the same Euclidean distance (78.3%). The recognition rate further jumps to 98% after we apply the multi-level subspace analysis algorithm. This clearly demonstrates that there are indeed a significant amount of information contained in the video sequence.

Next we compare the temporal synchronization and non-synchronization results in the two columns of Table 1. We again see a clear improvement of recognition accuracy by the A-V temporal synchronization approach for all the classification methods. Notice that although the improvement for the video classification using subspace analysis is only 1.7%, it reflects over 45% reduction of the recognition error rate, thus is more impressive than the other results.

Table 1. Comparison of recognition results on the gray level appearance features.

		A-V temporal Synchronization (%)	Non-Synchronization (%)
Still Image	Euclidean Distance	61.0	53.9
	Subspace Analysis	85.8	80.3
Video	Euclidean Distance	78.3	74.9
	Subspace Analysis	98.0	96.3

Now we look at the results on spatially synchronized local wavelet features, summarized in Table 2. As expected, all results are further improved. The comparison among different methods further confirms our observation in Table 1. Notice the final recognition accuracy of the experiment using all the three algorithms, temporal synchronization, spatial synchronization, and multi-level subspace analysis, is 99%. This is a very high accuracy considering that this is a cross session recognition. Finally, we compare our video recognition method with existing video based face recognition methods, the nearest frame method [8] and the mutual subspace method [8][11], in Table 3. Notice that the results for existing methods in Table 3 are computed from the A-V temporal synchronized video sequence, and our subspace analysis method is also applied to the nearest frame method. So they are already better than the original methods. We can still clearly see the significant improvement of our algorithms with only 5% to 10% of their error rates.

5. Conclusion

In this paper, we have developed an effective video-based face recognition algorithm. The algorithm takes full advantage of all the spatial-temporal information in

the video sequence. In order to overcome the processing speed and data size problems, the spatial and temporal frame synchronization algorithm and multilevel subspace analysis algorithm are developed. Experiments on the largest available face video database have shown that all the three techniques are effective in improving the recognition performance. Near perfect recognition results are achieved by the new algorithm. It is a significant improvement comparing to still image based method and existing video based method.

Table 2. Comparison of recognition results on the local wavelet features.

		A-V temporal Synchronization (%)	Non-Synchronization (%)
Still Image	Euclidean Distance	71.2	65.4
	Subspace Analysis	94.2	86.4
Video	Euclidean Distance	82.7	80.3
	Subspace Analysis	99.0	97.6

Table 3. Comparison of recognition results with existing video based methods.

Video-based methods	Recognition accuracy (%)
Mutual Subspace	79.3
Nearest frame using Euclidean distance	81.7
Nearest frame using subspace analysis	93.2
Multi-level Subspace using gray features	98.0
Multi-level Subspace using wavelet features	99.0

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region (Project no. CUHK 4190/01E and CUHK 4224/03E).

Reference

[1] V. Belhumeur, J. Hespanda, and D. Kiregeman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on PAMI,* Vol. 19, No. 7, pp. 711-720, July 1997.

[2] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition,* Vol. 33, pp. 1771-1782, 2000.

[3] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, pp.1090-1104, Oct. 2000.

[4] M. Turk and A. Pentland, "Face recognition using eigenfaces," *IEEE International Conference Computer Vision and Pattern Recognition*, pp. 586-591, 1991.

[5] L. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp.775-779, 1997.

[6] W. Zhao, R. Chellappa, and N. Nandhakumar, "Empirical performance analysis of linear discriminant classifiers," *Proceedings of CVPR,* pp. 164-169, 1998.

[7] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Matitre, "XM2VTSDB: The extended M2VTS database," *Second International Conference on AVBPA*, March 1999.

[8] S. Satoh, "Comparative evaluation of face sequence matching for content-based video access," *In Proceedings of IEEE International Conference on Automatic Face and Gesture*, Page(s): 163–168, 2000.

[9] V. Kruger and S. Zhou, "Exemplar-based face recognition from video," *In Proceedings of IEEE International Conference on Automatic Face and Gesture*, Page(s): 182 –187, 2002.

[10] G. Edwards, C. Taylor, and T. Cootes, "Improving identification performance by integrating evidence from sequences," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 486-491, 1999.

[11] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," *In Proceedings of IEEE International Conference on Automatic Face and Gesture*, Page(s): 318 –323, 1998.

[12] X. Wang and X. Tang, "Unified subspace analysis for face recognition," *Proceeding of IEEE International Conference on Computer Vision*, 2003.

[13] X. Tang, "Texture information in run-length matrices," *IEEE Transactions on Image Processing*, vol. 7, No. 11, pp. 1602-1609, Nov. 1998.

[14] T. F. Cootes, C. J. Edwards, and C. J. Taylor, "Active appearance models." IEEE Trans. on PAMI, Vol. 23, No. 6, pp. 681-685, June, 2001.