Video Based Face Recognition Using Multiple Classifiers

Xiaoou Tang and Zhifeng Li

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

Abstract

The main advantage of the video based face recognition 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, we develop a multiple classifiers fusion based video face recognition algorithm. The method preserves all the spatial-temporal information contained in a video sequence. A high recognition rate (98.6%) is achieved on the XM2VTS face video database.

1. Introduction

Face recognition has attracted a great deal of attention in recent years. Researchers have developed a number of promising algorithms including local feature analysis methods such as the Active Appearance Model (AAM) [2] and the elastic graph matching (EGM) method [12], and the appearance-based subspace methods such as the eigenface method [10], the LDA method [1][15], and the Bayesian algorithm [7]. However, all of these methods focus exclusively on image-based face recognition. One problem with the image-based method is that it is possible to use a pre-recorded face photo to pretend as a live subject. The second problem is that the image-based recognition accuracy is still too low to be used in some practical applications. In order to overcome these problems, video based face recognition has been proposed recently [3][5][9][14]. 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 in front of a live video camera may ensure the biometric data come from the actual user. 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 used, we may further increase the recognition accuracy.

Most researches on face recognition in video focus primarily on face detection and tracking in video. Once a face is detected in a video frame, the conventional image based face recognition technique will be used for single frame recognition. For recognition directly using video data, Satoh [9] matches two video sequences by selecting the pair of frames that are closest across the two videos. This is inherently still image-to-image matching. Methods in [3][5] use video sequence to train a statistical model face for matching. Even though the trained model is more robust than a model trained from a single image, the overall information contain in the model is still similar to a single image with the same feature dimension. This is similar to image-to-image matching with more training data. The mutual subspace method in [9][14] 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 much lower than other methods.

In this paper, we propose a multiclassifier-based video-to-video face recognition algorithm that takes full advantage of the complete spatial temporal information contained in a video sequence. We first use audio signal to align frames of similar images across the two video sequences so that they can be better matched. Then, for fast matching of the large video sequence, we use a multiple subspace classifiers fusion algorithm. Experiments on the largest standard video face database, the XM2VTS database [6], clearly demonstrate the effectiveness of the new algorithm.

2. Video-Based Face Recognition Using Multi-Classifiers

2.1 Video frame alignment

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 faces match with neutral faces and smile faces match with smile faces. Therefore, if we want to use video sequence for face recognition, it is important to align similar video frames in different video sequence. In order to accomplish this, we use a simple approach by utilizing information in the audio signal of the video. For example, in the XM2VTS database, 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. We locate the maximum point of each word and select the corresponding video frames. We can see different expressions when one read different words. 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. It is good enough for recognition purpose. The audio-guided method helps us to synchronize video sequence and select a number of distinctive frames for face recognition.

2.2 Multiple classifiers integration

After the video synchronization, 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 video sequence 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.

To overcome these problems, we develop a multiple classifiers based algorithm. We first use unified subspace analysis [11] classifier to process each individual video frame. Then all the frame-based classifiers are integrated using a fusion rule to determine the final classification. The detail algorithm is as follows.

1. Project each frame to its PCA subspace computed from the training set of the frame and adjust the PCA dimension to reduce noise.
2. For each frame, compute the whitened intrapersonal subspace using the within-class scatter matrix in the reduced PCA subspace and adjust the dimension of the whitened 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 above 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 for each frame.
5. Classify each frame using the discriminant feature vector computed in Step 4.
6. Combine all the frame-based classifiers using a fusion rule for final classification of the video sequence.

It has been shown that LDA can be implemented in three steps: PCA, within class whitening, and between class discriminant analysis. However, in each processing step, the subspace dimension is fixed at the maximum possible number. The difference between the traditional LDA and the unified subspace analysis [11] is that we allow the dimension in each step to change. This will not only help to reduce the feature dimension but can also help to remove more noisy features to improve the recognition performance.

Many methods on combining multiple classifiers have been proposed [4][13]. In this paper, we use two simple fusion rules in Step 6 to combine the frame-based classifiers: majority voting and sum rule.

Majority voting

Each classifier $C_k(x)$ assigns a class label to the input face data, $C_k(x) = i$. We represent this event as a binary function,

$$T_k(x \in X_i) = \begin{cases} 1, & C_k(x) = i \\ 0, & otherwise \end{cases}. \tag{1}$$

By a majority voting, the final class is chosen as,

$$\beta(x) = \arg\max_{X_i} \sum_{k=1}^{K} T_k\left(x \in X_i\right). \tag{2}$$

Sum rule

We assume that $P(X_i \mid C_k(x))$ is the probability that x belongs to X_i under the measure of the frame-based classifier $C_k(x)$. According to the sum rule, the class for the final decision is chosen as,

$$\beta(x) = \arg\max_{X_i} \sum_{k=1}^{K} P(X_i \mid C_k(x)) \tag{3}$$

$P(X_i \mid C_k(x))$ can be estimated from the output of the frame-based classifier. For the frame-based classifier $C_k(x)$, the center m_i of class X_i, and input face data x are projected to the discriminant vectors W_k,

$$w_k^i = W_k^T m_i \tag{4}$$

$$w_k^x = W_k^T x \tag{5}$$

$P(X_i \mid C_k(x))$ is estimated as

$$\hat{P}(X_i \mid C_k(x)) = \left(1 + \frac{\left(w_k^x\right)^T \left(w_k^i\right)}{\left\|w_k^x\right\| \cdot \left\|w_k^i\right\|} \right) / 2\,, \tag{6}$$

which has been mapped to [0,1].

3. Experiments

In this section, we conduct experiments on the XM2VTS face video database [6]. We select 294*4 video sequences of 294 distinct persons from the four different sessions. For the training data, we use 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 read two sequences, “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 guided video 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. 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 compare recognition results between still image based method and video based method. The results for both still image and video sequence are summarized in Table 1. The still images are either selected from the audio synchronized video sequence (A-V Synchronization case), or selected randomly from the non-synchronized video sequence (A-V Non-Synchronization case). For each individual frame, we compute the recognition accuracy. Then the average accuracy for all frames is shown in Table 1. We can see that the performance of using still image directly by Euclidean distance classification is very poor (62.3%). This baseline result 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 using video data. The recognition rate is improved to 98.6% using the voting rule for classifier integration. Figure 1 clearly illustrates the performance improvement. The results demonstrate that there is indeed significant amount of information contained in the video sequence.

Next, we compare the audio 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 synchronization approach for all the classification methods. All the methods reduce classification errors by 30% to 40% using AV synchronization.

Table 1. Comparison of recognition accuracies between still image based methods and video based methods.

		Non-Synchronization (%)	A-V Synchronization (%)
Still Image	Euclidean Distance	58.7	62.3
	Subspace Analysis	81.4	87.9
Video	Sum Rule	96.6	98.0
	Voting Rule	98.0	98.6

This is a very high accuracy considering that the testing data and gallery data are in different sessions. Finally, we compare our video recognition method with existing video based face recognition methods, the nearest frame method [9] and the mutual subspace method [9][14], in Table 2. Notice that the results for existing methods in Table 2 are computed from the A-V 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.

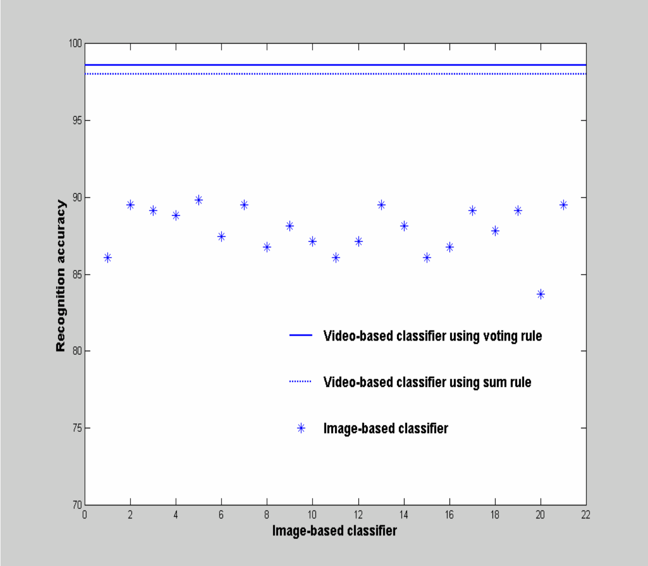

Figure 1. Comparison of video-based algorithms with individual image-based classifier based on the A-V data.

Table 2. Comparison of recognition accuracies with existing video based methods.

Video-based methods	Recognition Accuracy (%)
Mutual Subspace	79.3
Nearest frame using Euclidean distance	81.7
Nearest frame using LDA	90.9
Nearest frame using unified subspace analysis	93.2
Video-based classifier using sum rule	98.0
Video-based classifier using voting rule	98.6

4. 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, we propose a multiple classifiers fusion algorithm for video classification. Experiments on the largest available face video database have shown that the algorithm is 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.

5. 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).

6. 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] 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.

[3] 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.

[4] T. K. Ho, J. Hull, and S. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Trans. on PAMI*, Vol. 16, No.1, pp. 66-75, Jan. 1994.

[5] 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.

[6] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Matitre, “XM2VTSDB: The Extended M2VTS Database,” *Second International Conference on AVBPA*, March 1999.

[7] B. Moghaddam, T. Jebara, and A. Pentland, “Bayesian Face Recognition,” *Pattern Recognition,* Vol. 33, pp. 1771-1782, 2000.

[8] 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.

[9] 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.

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

[11] X. Wang and X. Tang, "Unified subspace analysis for face recognition," *In Proceedings of IEEE International Conference on Computer Vision,* pp. 679-686, 2003.

[12] 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.

[13] L. Xu, A. Krzyzak, and C. Y. Suen, "Method of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," *IEEE Trans. on System, Man, and Cybernetics*, Vol. 22, No. 3, 418-435, 1992.

[14] 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.

[15] W. Zhao, R. Chellappa, and N. Nandhakumar, “Empirical Performance Analysis of Linear Discriminant Classifiers,” *Proceedings of CVPR,* pp. 164-169, 1998.