Audio-Guided Video-Based Face Recognition

Xiaoou Tang, *Fellow, IEEE*, and Zhifeng Li, *Member, IEEE*

***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, multilevel discriminant subspace analysis, and multiclassifier integration for video sequence processing. An aligned video sequence for each person is first obtained by applying temporal and spatial synchronization, which effectively establishes the face correspondence using both audio and video information; then multilevel discriminant subspace analysis or multiclassifier integration is employed for further analysis based on the synchronized sequence. The method preserves most of the temporal-spatial information contained in a video sequence. Extensive experiments on the XM2VTS database clearly show the superiority of our new algorithms with near-perfect classification results (99.3%) obtained.**

***Index Terms*— Face recognition, subspace analysis, video processing.**

I. Introduction

AUTOMATIC face recognition is a challenging task in pattern recognition research. In recent years, a large number of methods have been developed [1], [2], [10], [11], [13], [17], [18], [20], [23], [27], [28], [31], [32]. Many of these methods and their combinations have shown promising recognition performance. However, 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 prerecorded 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 compared to other high-accuracy biometric technologies. To alleviate these problems, video-based face recognition has been proposed recently [4], [7]–[9], [12], [14]–[16], [21], [22], [30]. One of the major advantages of video-based face recognition is to prevent fraudulent system penetration by prerecorded 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 recognition directly using video data, Satoh [21] 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 [4], [7], [12], and [14] use a video sequence to train a statistical face model 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 limited. The mutual subspace method in [21] and [30] 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. Recently, Lee *et al*. [8], [9] try to model and recognize human faces in video sequences using probabilistic appearance manifolds. A limitation with the manifold learning algorithms is that they are based on modeling the characterization of "locality," i.e., it is more for feature representation than for classification. The method in [15] details a recognition system based on adaptive hidden Markov models (HMMs), each of which learns the temporal dynamics within a video sequence. By comparing likelihood scores yielded by the HMMs, the identity of a test video sequence is recognized with the highest score. Because learning temporal statistics during the recognition stage is very time consuming, the method is not practical for actual application.

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

Manuscript received June 28, 2006; revised April 19, 2007 and October 30, 2007. First version published May 12, 2009; current version published July 22, 2009. This work was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region (Project no. CUHK 4190/01E, CUHK 4224/03E, and CUHK 1/02C). This paper was recommended by Associate Editor C. Guillemot.

X. Tang is with the Department of Information Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong, China (e-mail: xtang@ie.cuhk.edu.hk).

Z. Li is with the Human–Computer Communications Laboratory, Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong (CUHK), Hong Kong, China (e-mail: zfli@se.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2009.2022694

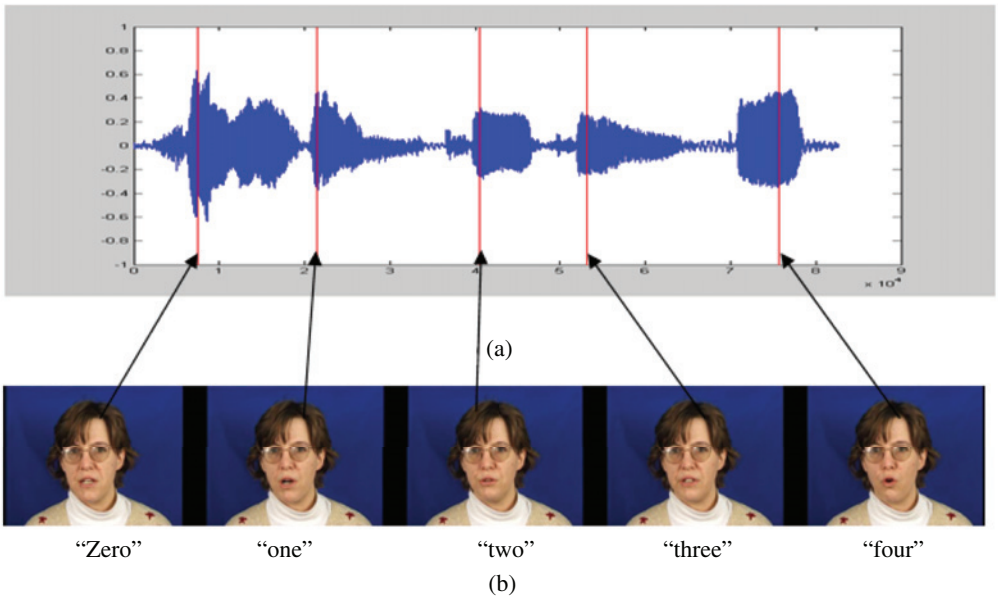


Fig. 1. Example video sequence and corresponding speech signal. (a) Speech signal and (b) face video frames.

First we develop a video frame temporal synchronization scheme. The idea is to align frames of similar images across the two video sequences so that they can be matched better. Given the large amount of data in video, we cannot afford to use a complicated algorithm for this purpose. We propose an efficient 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 [28]. Alignment of the fiducial points is critical for subspace methods to take advantage of the shape correlation across different face images. More accurate face alignment methods may be used for this purpose if computational cost is tolerable. Finally, for matching of the large spatial and temporal synchronized video sequence effectively and efficiently, we develop a multilevel discriminant subspace analysis algorithm and a multiclassifier integration algorithm. Extensive experiments on the largest standard video face database, i.e., the XM2VTS database [19], clearly show the feasibility and effectiveness of our new algorithms with high recognition accuracy achieved.

II. 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. Otherwise, there is no additional information to extract from such video sequences. However, for videos of varying frame contents, a simple frame-by-frame matching of the two video sequences may not lead to significant performance improvement 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 sequences. 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 videos. 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 contain video sequences for 295 people. There are totally 295×4 video sequences of 295 distinct persons for experiments. For each person, several video sequences of 20 s 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 facial deformations when one reads different words. Of course, more sophisticated speech recognition techniques can also be used to improve the result with added computational cost. We found our simple approach already 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 faces will move and change. We call this step spatial synchronization. Alignment of the fiducial points is

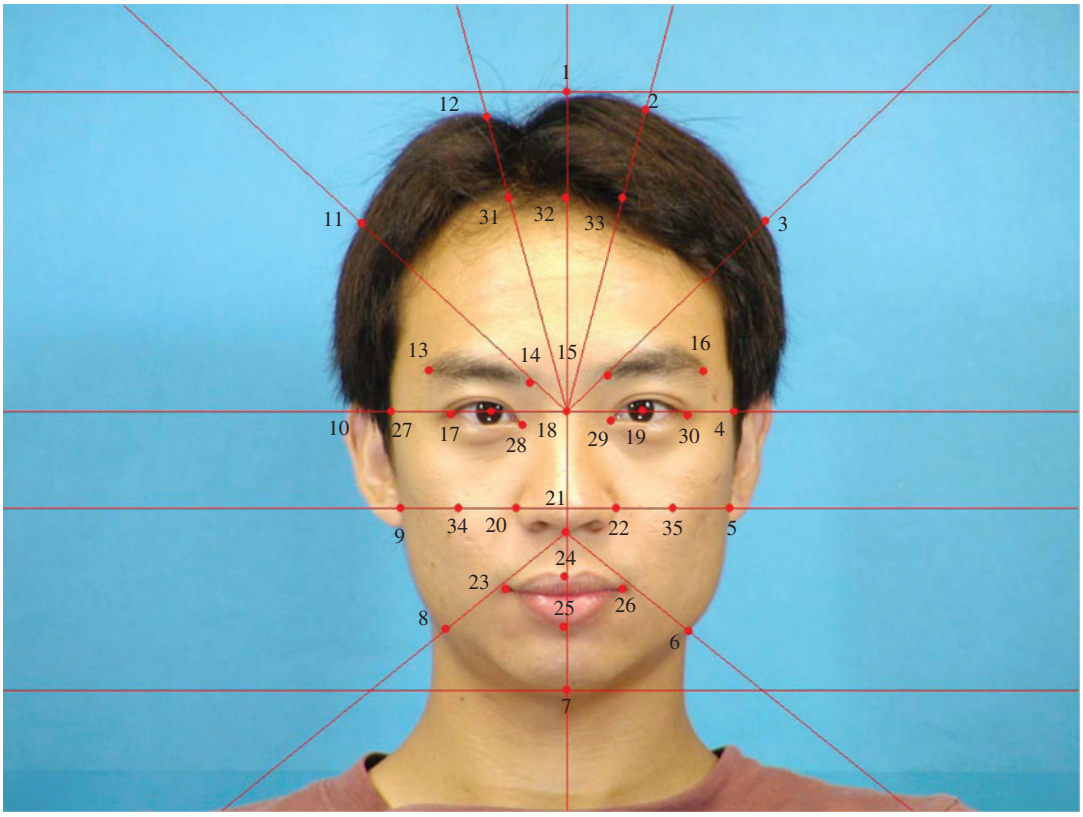


Fig. 2. Fiducial graph model.

critical for subspace methods to take advantage of the shape correlation across different human faces. We design a fiducial grid model with 35 fiducial points over a face image as shown in Fig. 2. This model is different from that of other research. We first define a set of anchor points over salient locations such as nose tip and mouth corners on the face images so that they can be located easily and accurately. Then, other fiducial points can be derived from these anchor points. For example, in Fig. 2, point 8 can be located through intersection of the face contour with a line anchored by the nose tip (point 21) and the mid-point of points 7 and 9. We use the Gabor wavelet features [28] to allocate key fiducial points for the spatial synchronization [24]. We choose five scales and eight orientations in our study. Thus a set of 40 Gabor features can be obtained for each local fiducial point. The face image is represented by a large Gabor feature vector combining 35 local vectors. It is certainly possible to use active appearance model to improve the alignment accuracy. Unfortunately, the computational cost is too high for our application.

III. Multilevel Discriminant Subspace Analysis

A. Methodology

After the spatial and temporal synchronization, we finally have an aligned video sequence that is composed of 21 large Gabor feature vectors 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 sequence as a single large feature vector and conduct regular subspace analysis to extract features. Although this feature-level fusion approach utilizes 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 41×27 for each video sequence, thus the feature dimension is 23 247. Direct subspace analysis on such a large vector is too costly. Traditionally, for image-based face recognition, when the sample number is small, researchers have used the dominant eigenvector estimation method to compute the eigenvectors in the dual space [23]. However, this method limits the number of samples one can use for training. In addition, for video-based recognition, the limit on the sample number is even stricter due to extremely high computer memory size requirement. Since to compute the eigenvectors in the dual-space, we have to construct the training sample matrix containing all the samples of the huge video feature vectors. In our experiment, the matrix size is $3 \times 295 \times 23\,247$. Manipulating such a huge matrix would put extremely high demand on computer memory. We failed to carry out this operation in our computer with 2 GB memory. Finally, a more serious problem is the overfitting problem because of the small sample size vs. large feature dimension for discriminant subspace analysis algorithms. This has always been a problem for image-based face recognition. It gets even worse for the video-based face recognition.

In order to overcome these problems, inspired by the multilevel dominant eigenvector estimation method [26], we

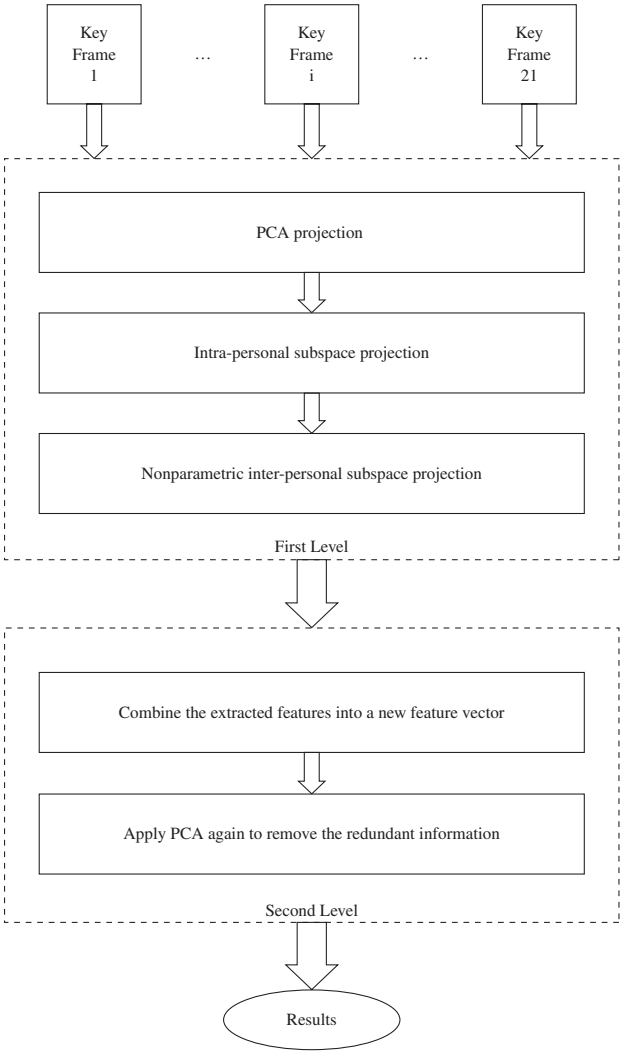


Fig. 3. Illustration of multilevel discriminant subspace analysis.

develop a multilevel discriminant subspace analysis algorithm, as illustrated in Fig. 3. We first break the video sequence into slices, with features from each frame as a slice. Then we perform discriminant subspace analysis on each feature slice. The extracted discriminant features from each slice are then concatenated to form a new feature vector. We then apply principal component analysis (PCA) to the new feature vector to remove redundant information among the feature slices to extract the final feature vector. The details of our algorithm are as follows.

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

1) project each frame F to its PCA subspace using the transform matrix W_1 computed from the training set for dimensionality reduction
$$F_1 = W_1^T F; \tag{1}$$
2) compute the within-class subspace using the within-class scatter matrix in the reduced PCA subspace and then adjust the dimension of intrapersonal subspace to better reduce the intrapersonal variation;
3) project each frame (after PCA transform) onto the intrapersonal subspace, and then normaize the projections by intrapersonal eigenvalues to compute the whitened feature vectors
$$F_2 = W_2^T F_1 \tag{2}$$
where W_2 is the normalized projection matrix;
4) calculate the nonparametric between-class scatter matrix [13] based on the whitened feature vectors;
5) compute the final discriminant feature vector using PCA transform matrix W_3 based on the calculated nonparametric scatter matrix in step 4
$$F_3 = W_3^T F_2. \tag{3}$$

In the second level of subspace analysis:

1) combine the extracted discriminant feature vectors from each slice into a new feature vector F_{new};
2) apply PCA on the new feature vector to remove redundant information as much as possible. The dominant features with large eigenvalues are selected to form the projection matrix W_4 to compute the final feature vector for recognition
$$F_{\text{final}} = W_4^T F_{\text{new}}. \tag{4}$$

LDA [1], [32] is a popular face subspace analysis technique. 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. Similar to the unified subspace analysis method [27], the advantage of our new algorithm over the traditional LDA is that we allow the dimension in each step to change, as shown in the first five steps of the first level of the above algorithm. This will not only help to reduce the feature dimension but also help to remove more noisy features to improve the recognition performance. Next we discuss in more detail the steps of the algorithm.

B. Nonparametric Subspace Analysis

In LDA, another serious problem stems from the parametric nature of the scatter matrix. The construction of the scatter matrix in LDA is based on the underlying assumption that the samples in each class share the same Gaussian distribution. Therefore, LDA performs well under Gaussian class distributions, but not under non-Gaussian distributions. The difference between LDA and nonparametric subspace analysis lies in the definition of the between-class scatter matrix. In LDA, the between-class scatter matrix S_b is defined in a parametric form

$$S_b = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{5}$$

where μ_i denotes the mean of the class C_i, and N_i denotes the number of samples in class C_i. Such a parametric formulation leads to several disadvantages. First, LDA is based on the assumption that the discrimination information is equal for all classes. Therefore, its performance notably degrades in case of non-Gaussian distributions. Second, due to the presence of noise, the between-class matrix S_b in LDA cannot capture the information of the boundary structure effectively. In order to overcome these problems, the proposed multilevel

discriminant subspace analysis algorithm adopts a new feature extraction technique called multiclass nonparametric subspace analysis [13], where a nonparametric between-class scatter matrix S_b^N is defined as

$$S_b^N = \sum_{i=1}^{c} \sum_{\substack{j=1 \\ j \neq i}}^{c} \sum_{l=1}^{N_i} w(i,j,l) \left(x_l^i - m_j\left(x_l^i\right)\right)\left(x_l^i - m_j\left(x_l^i\right)\right)^T \tag{6}$$

where x_l^i denotes the lth face vector of class i, and $m_j(x_l^i)$ is the local KNN mean, defined by

$$m_j\left(x_l^i\right) = \frac{1}{k} \sum_{p=1}^{k} NN_p\left(x_l^i, j\right) \tag{7}$$

where $NN_p(x_l^i, j)$ is the pth nearest neighbor from class j to the face vector x_l^i. $w(i, j, l)$ is the weighting function, defined as

$$w(i,j,l) = \frac{\min\left\{d^\alpha\left(x_l^i, NN_k\left(x_l^i, i\right)\right), d^\alpha\left(x_l^i, NN_k\left(x_l^i, j\right)\right)\right\}}{d^\alpha\left(x_l^i, NN_k\left(x_l^i, i\right)\right) + d^\alpha\left(x_l^i, NN_k\left(x_l^i, j\right)\right)} \tag{8}$$

where α is a parameter ranging from zero to infinity, which controls the changing speed of the weight w.r.t the distance ratio. $d(v_1, v_2)$ is the Euclidean distance between two vectors v_1 and v_2. The weighting function has the property that, for samples near the classification boundary, it approaches 0.5 and drops off to zero if the samples are far away from the classification boundary. By using such a weighting function, the boundary information contained in the training set is emphasized. In (6), all the training samples are used to calculate the nonparametric between-class scatter matrix instead of merely using the class centers. In addition, the weighting function defined in (8) can emphasize the boundary structure information, which can help to improve the classification accuracy.

In the second-level subspace analysis, we only use PCA instead of the discriminant 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 steps 4 and 5 of the first level. Repeating them will not add new information. However, there is a significant amount of information overlap between different slices since the frames are still quite similar to each other even with expression changes. PCA is needed to remove redundant features.

C. Discussions on Multilevel Subspace Method

In this section, we show that the multilevel discriminant subspace analysis does not lose much information compared to the original subspace analysis. Since steps 2 to 5 of the first level aim to remove intrapersonal variations which contain only unwanted information, we do not need to consider them when analyzing information loss in the multilevel discriminant subspace analysis algorithm. So we only need to focus on the two PCA steps; step 1 of the first level and step 2 of the second level. Thus the analysis is similar to the one in [26]. Here we show more detailed proof.

To compute PCA, we first form an n by m sample matrix

$$A = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_m(1) \\ x_1(2) & x_2(2) & \dots & x_m(2) \\ \dots & \dots & \dots & \dots \\ x_1(n) & x_2(n) & \dots & x_m(n) \end{bmatrix} \tag{9}$$

where x_i is a 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 (as shown in the equation at the bottom of the page) 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 show 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

$$A = \begin{bmatrix} B_1 \left\{ \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_m(1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(k) & x_2(k) & \cdots & x_m(k) \end{bmatrix} \right\} \\ B_2 \left\{ \begin{bmatrix} x_1(k+1) & x_2(k+1) & \cdots & x_m(k+1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(2k) & x_2(2k) & \cdots & x_m(2k) \end{bmatrix} \right\} \\ B_3 \left\{ \begin{bmatrix} x_1((g-1)k+1) & x_2((g-1)k+1) & \cdots & x_m((g-1)k+1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(n) & x_2(n) & \cdots & x_m(n) \end{bmatrix} \right\} \end{bmatrix}$$

its covariance matrix are

$$A = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \tag{10}$$

$$W = AA^T = \begin{bmatrix} B_1 B_1^T & B_1 B_2^T \\ B_2 B_1^T & B_2 B_2^T \end{bmatrix} = \begin{bmatrix} W_1 & W_{12} \\ W_{21} & W_2 \end{bmatrix}. \tag{11}$$

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

$$T_1^T W_1 T_1 = \Lambda_1 \tag{12}$$

$$T_2^T W_2 T_2 = \Lambda_2 \tag{13}$$

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

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}. \tag{14}$$

T is also an orthogonal matrix, since

$$T^T T = \begin{bmatrix} T_1^T T_1 & 0 \\ 0 & T_2^T T_2 \end{bmatrix} = I. \tag{15}$$

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

$$\begin{aligned} W_r = T^T W T &= \begin{bmatrix} \Lambda_1 & T_1^T W_{12} T_2 \\ T_2^T W_{21} T_1 & \Lambda_2 \end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} \Lambda_{1b} & 0 \\ 0 & \Lambda_{1s} \end{bmatrix} & \begin{bmatrix} C_{bb} & C_{bs} \\ C_{sb} & C_{ss} \end{bmatrix}^T \\ \begin{bmatrix} C_{bb} & C_{bs} \\ C_{sb} & C_{ss} \end{bmatrix} & \begin{bmatrix} \Lambda_{2b} & 0 \\ 0 & \Lambda_{2s} \end{bmatrix} \end{bmatrix} \end{aligned} \tag{16}$$

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 (16) to discuss the impact on W by keeping only the first few dominant eigenvalues in each group. In (16), 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

$$W_d = \begin{bmatrix} \Lambda_{1b} & C_{bb}^T \\ C_{bb} & \Lambda_{2b} \end{bmatrix}. \tag{17}$$

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 (16) 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 are partially correlated, the correlated part should be mostly signal, since noise parts of the variables B_1 and B_2 do not 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 small relative to other terms in W_r, and so not much information is lost 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 simplified multilevel subspace method are indeed a close approximation of the eigenvalues computed from W of the standard PCA method. The final eigenvectors for the two-group case is

$$V = \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} V_{PCA} \tag{18}$$

where V_{PCA} is the projection matrix of the final PCA step.

We now use a simple experiment to verify the above analysis. We apply the PCA and the multilevel PCA subspace separately on the XM2VTS face image database [19]. For the training data, we select 295×3 images of 295 people from the first three sessions. The gallery set is composed of 295 images of 295 people from the first session. The probe set is composed of 295 images of 295 people from the fourth session.

For the multilevel PCA subspace, the feature vector with length 1107 is broken into nine small feature vectors of length 121 each. Then, the first $k = 60$, 40, 20, and 10 dominant features in each group are selected to form the new feature vector of length $9 \times k$, from which the final dominant eigenvalues and eigenvectors are computed. Fig. 4(a) shows the results of the top 20 eigenvalues of the standard PCA and the multilevel PCA subspace with the three values of k. We see that when 60, 40, or 20 features are kept after the first-step eigenvalue computation, the final eigenvalues are almost the same as the standard PCA. When only ten features are kept, the first 15 eigenvalues are still similar to PCA; the remaining eigenvalues start to lose a very small amount of energy. However, this does not affect the final classification results much. Fig. 4(b) shows the classification accuracies plotted against the number of features used. All four groups of results overlap with each other almost completely.

Using such a method, we are no longer limited by either the size of the video sequence or the number of training

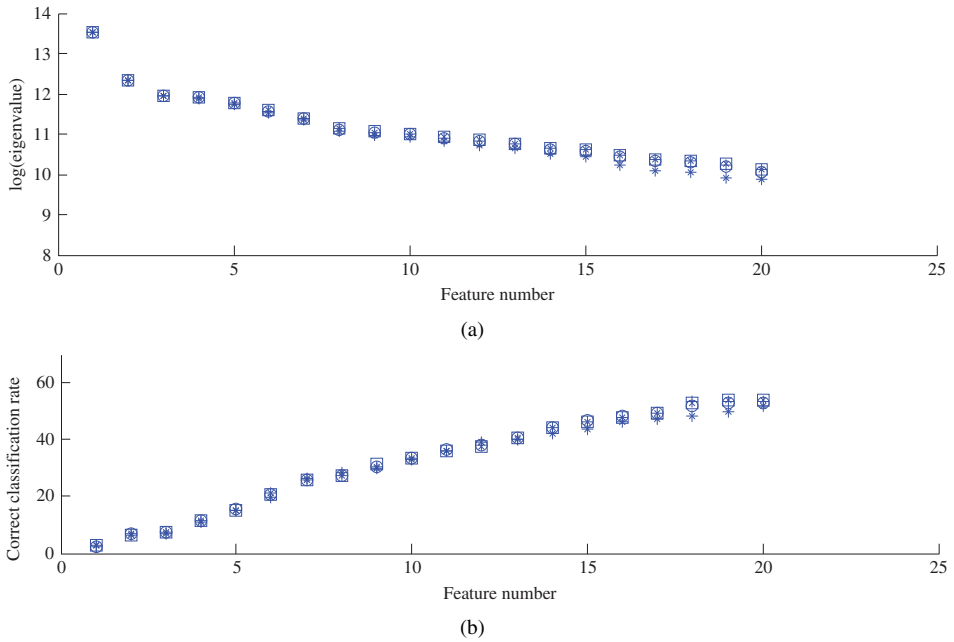


Fig. 4. Comparison of the PCA and multilevel PCA subspace: (a) plot of the top 20 largest eigenvalues and (b) plot of the correct classification rate against number of features used. In both plots, the square symbol is for the original PCA and the other three are for multilevel PCA subspace: "x" for $k = 60$; "+" for $k = 40$; "O" for $k = 20$ and star symbol for $k = 10$.

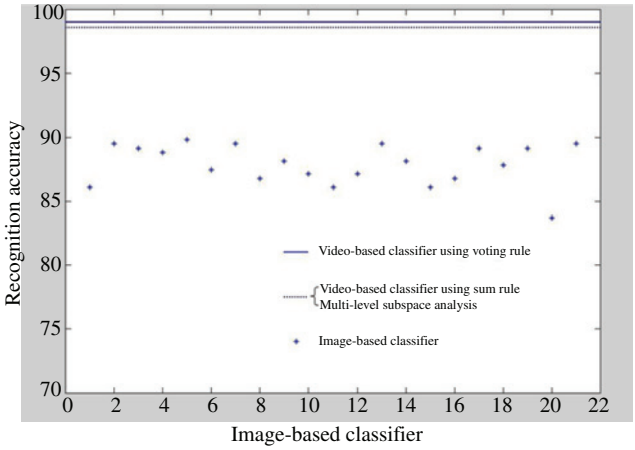


Fig. 5. Comparison of video-based algorithms with individual image-based classifier on the gray level appearance features.

samples. The multilevel discriminant subspace analysis algorithm has several desirable advantages. First, the first level of the algorithm is composed of several parallel subspace analysis classifiers with each one dealing with a portion of the data; thus efficient parallel computation is possible. Secondly, in the second level, the algorithm significantly removes redundant information producing a compact feature vector. Third, it embeds the nonparametric subspace analysis technique, which has been shown to outperform the traditional LDA. Accordingly, it can help to enhance the recognition performance with more accurate final classifier constructed.

IV. Multiple Classifiers Integration

The second-level subspace analysis in the multilevel discriminant subspace analysis can be considered as a feature-level fusion strategy to combine information from different frames. In fact, we can also replace it with a decision-level fusion strategy by using multiple classifiers integration. In the first level, we still use the nonparametric subspace analysis (NSA) based classifier [13] to process each individual video frame. Then all the frame-based classifiers are integrated using a fusion rule to determine the final classification. The detailed algorithm is as follows.

The first-level subspace analysis remains the same. The second level of processing is changed as follows.

1) Classify each frame using the discriminant feature vectors computed in steps 4 and 5, respectively.
2) Combine all the frame-based classifiers using a fusion rule for final classification of the video sequence.

Many methods on combining multiple classifiers have been proposed [6], [29]. They can all be used in our algorithm. Among the existing fusion rules, majority voting and sum rule are among the most representative ones. They have been shown to be both efficient and effective compared with most other fusion methods [6], [29]. In our previous work [25], we have used the two fusion rules and achieved very encouraging results.

A. 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

$$T_k(x \in X_i) = \begin{cases} 1, & C_k(x) = i \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

By a majority voting, the final class is chosen as

$$\beta(x) = \arg\max_{X_i} \sum_{k=1}^{K} T_k(x \in X_i) \tag{20}$$

where K denotes the number of individual classifiers.

TABLE I
COMPARISON OF RECOGNITION RESULTS ON THE GRAY LEVEL APPEARANCE FEATURES

Recognition methods		A-V temporal synchronization recognition rate (%)	Non-synchronization recognition rate (%)
Still image	Euclidean distance	61.0	53.9
	Subspace analysis	85.8	80.3
Video	Euclidean distance	78.3	74.9
	Multilevel discriminant subspace analysis	98.6	96.9
	Multiclassifier voting rule	99.0	97.6
	Multiclassifier sum rule	98.6	96.9

TABLE II
COMPARISON OF RECOGNITION RESULTS ON THE LOCAL WAVELET FEATURES

Recognition methods		A-V temporal synchronization recognition rate (%)	Non-synchronization recognition rate (%)
Still image	Euclidean distance	71.2	65.4
	Subspace analysis	94.2	86.4
Video	Euclidean distance	82.7	80.3
	Multilevel discriminant subspace analysis	99.3	98.0
	Multiclassifier voting rule	99.3	98.3
	Multiclassifier sum rule	99.3	98.0

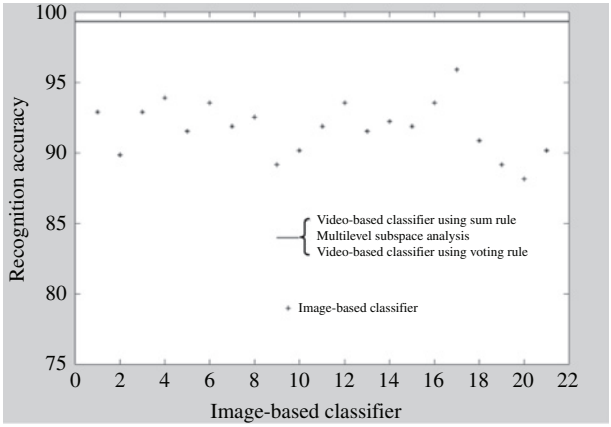


Fig. 6. Comparison of video-based algorithms with individual image-based classifier based on the local wavelet features.

B. Sum Rule

We assume that $P(X_i|C_k(x))$ 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

$$\beta(x) = \arg\max_{X_i} \sum_{k=1}^{K} P(X_i|C_k(x)). \tag{21}$$

$P(X_i|C_k(x))$ 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

$$w_k^i = W_k^T m_i \tag{22}$$

$$w_k^x = W_k^T x. \tag{23}$$

$P(X_i|C_k(x))$ is estimated as

$$\hat{P}(X_i|C_k(x)) = \left(1 + \frac{(w_k^x)^T (w_k^i)}{\|w_k^x\| \cdot \|w_k^i\|}\right) \Big/ 2 \tag{24}$$

which has been mapped to [0, 1].

Compared with the multilevel discriminant subspace analysis algorithm, the multiple classifiers integration algorithm can provide more flexibility. More sophisticated fusion methods can be used in the second level to further boost the recognition performance in the future.

V. EXPERIMENTS

In this section, we conduct extensive experiments on the XM2VTS database [19]. There are many different still image-based face datasets available, but for video-based face recognition, the XM2VTS we use is the only publicly available large standard dataset containing 1180 video sequences from 295 persons. In our experiments, we use all the 295×4 video sequences of 295 persons from the four different sessions. For the training data, we select the 295×3 video sequences of the first three sessions. The gallery set is composed of the 295 video sequences of the first session. The probe set is composed of the 295 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 with each one having 21 video frames. For the A-V synchronization data, each frame corresponds to the waveform peak of a digit in the two recited sentences "0 1 2 3 4 5 6 7 8 9" and "5 0 6 9 2 8 1 3 7 4." An additional frame is located at the midpoint

TABLE III
COMPARISON OF RECOGNITION RESULTS WITH EXISTING VIDEO BASED METHODS (ALL METHODS ARE BASED ON A-V SYNCHRONIZED DATA)

Video-based methods	Recognition rate (%)	
	Gray level features	Wavelet features
Mutual subspace	79.3	N/A
Nearest frame using Euclidean distance	81.7	N/A
Nearest frame using LDA	90.9	N/A
Multiple LDA classifiers using sum rule	95.6	97.0
Multiple LDA classifiers using voting rule	95.9	97.3
Multilevel discriminant subspace	98.6	99.3
Video-based sum rule multiclassifier	98.6	99.3
Video-based voting rule multiclassifier	99.0	99.3

of the end of the first sentence and the start of the second sentence. For the A-V non-synchronization data, the 21 frames are randomly selected without using the audio information.

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 I. The still image is either selected from the first frame extracted by using the temporal synchronization (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 a still image directly by Euclidean distance classification is very poor (61.0% recognition rate). This baseline result actually reflects the difficulty of the database. In general, if the probe image and the gallery image are from different sessions in face recognition experiments, 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% recognition rate), where the extracted 21 video frames span a video sequence for Euclidean distance classification. The recognition error rate is drastically reduced to 1.4% or 1% after we apply the multilevel discriminant subspace analysis algorithm or multiclassifier integration on the video sequence. This clearly demonstrates that there is indeed significant amount of information contained in the video sequence and our new algorithms can utilize the additional information contained in the video sequence to improve the recognition accuracy. Fig. 5 clearly illustrates the performance improvement using the multilevel discriminant subspace analysis or multiclassifier integration over the individual image-based classifier on the gray level appearance features.

Next we compare the temporal synchronization and non-synchronization results in the two columns of Table I. We again see a clear improvement of recognition accuracy by the A-V temporal synchronization approach for all the classification methods, which clearly demonstrates the effectiveness and feasibility of our temporal synchronization method. Notice that although the difference between A-V synchronization and non-sychronization for the multiclassifier voting rule that achieves the best performance is only 1.4 percentage point, it reflects about 60% reduction of the recognition error rate.

Now we look at the results on spatially synchronized local wavelet features, reported in Table II and Fig. 6. As expected, all results are further improved. In addition, the comparison among different methods further confirms our observations in Table I and Fig. 5. Especially, notice the final recognition accuracy of the experiment using all the three algorithms: temporal synchronization, spatial synchronization, and multilevel discriminant subspace analysis (or multiclassifier), is 99.3%. This is a very high accuracy considering that this is cross-session recognition.

Finally, we compare our video recognition method with existing video-based face recognition methods, the nearest frame method [21], and the mutual subspace method [21], [30]. The comparative results are reported in Table III. Notice that the results for existing methods in Table III are computed from the A-V temporal synchronized video sequence, so they are already better than the original methods. We can still clearly see the significant improvement of our algorithms with only 5–10% error rates of traditional methods. In Table III we also combine traditional LDA with multiple classifiers based on the synchronized data. They give better results than all traditional methods, which again demonstrate the effectiveness of our video frame synchronization method. However, they are much less accurate than our best algorithms which embed the new subspace analysis technique. This clearly shows the superiority our new subspace analysis over the traditional LDA algorithm.

VI. CONCLUSION AND DISCUSSION

In this paper, we have developed a new video-based face recognition framework. The framework takes advantage of the temporal–spatial information in the video sequence. In order to overcome the processing speed and data size problems and extract the most discriminant features at the same time, the spatial and temporal frame synchronization algorithm, multilevel subspace analysis algorithm, and multiclassifier integration algorithm have been developed and incorporated into the framework. Experiments on the largest available face video database have shown that all the new techniques are effective in improving the recognition performance. Near-perfect recognition results are achieved on the test data by the new algorithms. It is a significant improvement compared to still image-based methods and existing video-based methods.

There are several directions for future work. An obvious improvement is to extend the audio-guided method to include more speech information. For example, speaker verification based on the user's voice can be incorporated into this framework to achieve better performance. It is also worthwhile to

use more sophisticated video analysis techniques to further explore the rich dynamical behaviors of the face expression and head movements. Finally, the weakness of the audio-guided method is that it is vulnerable to data perturbation. For example, if a person reads the digit sequence "0, 1, 2, . . . , 9" in a wrong order, skips one digit, or repeats one digit, then the frame synchronization may fail. A method more tolerant to these types of errors should be studied in future work.

REFERENCES

[1] P. Belhumeur, J. Hespanda, and D. Kiregeman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] L. Chen, H. Liao, J. Lin, M. Ko, and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[3] T. F. Cootes, C. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[4] G. Edwards, C. Taylor, and T. Cootes, "Improving identification performance by integrating evidence from sequences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition (CVPR)*, 1999, pp. 486–491.

[5] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[6] T. K. Ho, J. Hull, and S. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 1, pp. 66–75, Jan. 1994.

[7] V. Kruger and S. Zhou, "Exemplar-based face recognition from video," in *Proc. IEEE Int. Conf. Automat. Face Gesture*, 2002, pp. 182–187.

[8] K. Lee, J. Ho, M. Yang, D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2003, pp. 313–320.

[9] K. Lee, D. Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2005, pp. 852–859.

[10] Z. Li and X. Tang, "Using support vector machines to enhance the performance of bayesian face recognition," *IEEE Trans. Inform. Forensics Security (TIFS)*, vol. 2, no. 2, pp. 174–180, Jun. 2007.

[11] Z. Li and X. Tang, "Bayesian face recognition using support vector machine and face clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, vol. 2. Jun. 2004, pp. II-374–II-380.

[12] B. Li, R. Chellappa, Q. Zheng, and S. Der, "Model-based temporal object verification using video," *IEEE Trans. Image Process.*, vol. 10, no. 6, pp. 897–908, Jun. 2001.

[13] Z. Li, W. Liu, D. Lin, and X. Tang, "Nonparametric subspace analysis for face recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2005, pp. 961–966.

[14] B. Li, and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans Image Process.*, vol. 11, no. 5, pp. 530–544, May 2002.

[15] X. Liu, and T. Cheng, "Video-based face recognition using adaptive hidden Markov models," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, vol. 1. Jun. 2003, pp. 340–345.

[16] L. Liu, Y. Wang, and T. Tan, "Online Appearance Model Learning for Video-Based Face Recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2007, pp. 1–7.

[17] C. Liu, "Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 725–737, May 2006.

[18] M. Loog and R. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 732–739, Jun. 2004.

[19] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Matitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio-Video-Based Biometric Person Authentication (AVBPA)*, Washington, D.C., Mar. 1999, pp. 72–77.

[20] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, 2000.

[21] S. Satoh, "Comparative evaluation of face sequence matching for content-based video access," in *Proc. IEEE Int. Conf. Automat. Face Gesture*, 2000, pp. 163–168.

[22] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, "Video-based face recognition on real-world data," in *Proc. Int. Conf. Comput. Vision (ICCV)*, 2007, pp. 1–8.

[23] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 1991, pp. 586–591.

[24] X. Tang and Z. Li, "Frame synchronization and multi-level subspace analysis for video based face recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, vol. 2. Jun. 2004, pp. II-902–II-907.

[25] X. Tang and Z. Li, "Video based face recognition using multiple classifiers," in *Proc. 6th Int. Conf. Automat. Face Gesture Recognition (FG)*, May 2004, pp. 345–349.

[26] X. Tang, "Texture information in run-length matrices," *IEEE Trans. Image Process.*, vol. 7, no. 11, pp. 1602–1609, Nov. 1998.

[27] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.

[28] L. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.

[29] L. Xu, A. Krzyzak, and C. Y. Suen, "Method of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 3, pp. 418–435, Jun. 1992.

[30] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Proc. IEEE Int. Conf. Automat. Face Gesture*, 1998, pp. 318–323.

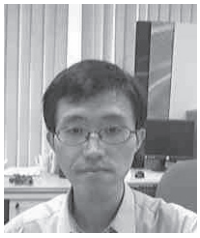
[31] J. Yang, D. Zhang, J. Yang, and B. Niu, "Globally maximizing, locally minimizing: Unsupervised discriminant projection with application to face and palm biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 650–664, Apr. 2007.

[32] W. Zhao, R. Chellappa, and N. Nandhakumar, "Empirical performance analysis of linear discriminant classifiers," in *Proc. IEEE Comput. Vision Pattern Recognition (CVPR)*, 1998, pp. 164–169.

Xiaoou Tang (S'93–M'96–SM'02–F'09) received the B.S. degree from the University of Science and Technology of China, Hefei, in 1990, the M.S. degree from the University of Rochester, Rochester, NY, in 1991, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1996.

He is a Professor with the Department of Information Engineering, The Chinese University of Hong Kong. He worked as the Group Manager of the Visual Computing Group, Microsoft Research Asia, from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing.

Dr. Tang has received the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition 2009. He is a program chair of the IEEE International Conference on Computer Vision 2009 and an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal of Computer Vision*.



Zhifeng Li (M'06) received the B.S. degree from the University of Science and Technology of China, Hefei, and the M.Phil and Ph.D. degrees in information engineering from the Chinese University of Hong Kong (CUHK), Hong Kong, China.

Currently, he is a Postdoctoral Fellow with the Department of Systems Engineering and Engineering Management, CUHK. His research interests include computer vision, pattern recognition, and multimodal biometrics.

Dr. Li has been a Program Committee Member for several international conferences such as CVPR, ICCV, ECCV, etc. He is a reviewer for a number of journals and conferences such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, European Conference on Computer Vision, International Conference on Multimedia and Expo, Asian Conference on Computer Vision, etc.

Dr. Li is a member of IEEE Computer Society.