International Journal of Computer Vision 70(1), 91–104, 2006

DOI: 10.1007/s11263-006-8098-z

Random Sampling for Subspace Face Recognition

XIAOGANG WANG AND XIAOOU TANG
Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

Received February 24, 2005; Revised November 27, 2005; Accepted January 4, 2006

First online version published in May, 2006

Abstract. Subspace face recognition often suffers from two problems: (1) the training sample set is small compared with the high dimensional feature vector; (2) the performance is sensitive to the subspace dimension. Instead of pursuing a single optimal subspace, we develop an ensemble learning framework based on random sampling on all three key components of a classification system: the feature space, training samples, and subspace parameters. Fisherface and Null Space LDA (N-LDA) are two conventional approaches to address the small sample size problem. But in many cases, these LDA classifiers are overfitted to the training set and discard some useful discriminative information. By analyzing different overfitting problems for the two kinds of LDA classifiers, we use random subspace and bagging to improve them respectively. By random sampling on feature vectors and training samples, multiple stabilized Fisherface and N-LDA classifiers are constructed and the two groups of complementary classifiers are integrated using a fusion rule, so nearly all the discriminative information is preserved. In addition, we further apply random sampling on parameter selection in order to overcome the difficulty of selecting optimal parameters in our algorithms. Then, we use the developed random sampling framework for the integration of multiple features. A robust random sampling face recognition system integrating shape, texture, and Gabor responses is finally constructed.

Keywords: random subspace method, bagging, LDA, face recognition, subspace analysis

1. Introduction

Subspace methods for face recognition have been extensively studied in recent years (Turk and Pentland, 1991; Belhumeur et al., 1997; Chen et al., 2000; Moghaddam et al., 2000; Wang and Tang, 2004a, 2004b, 2004c). Although having achieved great success, they still suffer from two problems: (1) the training sample set is small compared with the high dimensional feature vector; (2) the performance of subspace methods is sensitive to the subspace dimension. In this paper, we focus on Linear Discriminant Analysis (LDA), since it is a popular and widely studied subspace method in face recognition. But the conclusion can be extended to other subspace methods.

According to the Fisher criteria, LDA determines a set of projection vectors maximizing the between-class scatter matrix and minimizing the within-class scatter matrix in the projective feature space. But when dealing with the high dimensional face data, LDA often suffers from the small sample size problem. Since usually there are only a few samples in each face class for training, the within-class scatter matrix is not well estimated and may become singular. So the LDA classifier is often biased and sensitive to slight changes of the training set.

To address this problem, a two-stage PCA+LDA approach, i.e. Fisherface (Belhumeur et al., 1997) is proposed. Using PCA, the high dimensional face data is projected to a low dimensional feature space and then

LDA is performed in the low dimensional PCA subspace. Usually, the eigenfaces with small eigenvalues are removed in the PCA subspace. Since they may also encode some information helpful for recognition, their removal may introduce a loss of discriminant information. To construct a stable LDA classifier, the PCA subspace dimension depends on the training set size. When the PCA subspace dimension is relatively high, the constructed LDA classifier is often biased and unstable. The projection vectors may be greatly changed by the slight disturbance of noise on the training set. So when the training set is small, some discriminative information has to be discarded in order to construct a stable LDA classifier.

Chen et al. (2000) suggested that the null space spanned by the eigenvectors of the within-class scatter matrix with zero eigenvalues contains the most discriminative information. A LDA method in the null space of the within-class scatter matrix was proposed. It chose the projection vectors maximizing the between-class scatter matrix with the constraint that the within-class scatter matrix is zero. However, as explained in Chen et al. (2000), with the existence of noise, when the training sample number is large, the null space of the within-class scatter matrix becomes small, so much discriminative information outside this null space will be lost. The constructed classifier may also be over tuned to the training set.

In our previous work (Wang and Tang, 2004), based on a unified framework for subspace analysis, it is shown that LDA can be performed by three steps. Both Bayesian subspace analysis and PCA can be viewed as intermediate steps of LDA. It is shown that the subspace dimensions of the PCA subspace, the intrapersonal subspace, and the LDA subspace in the three steps can significantly affect the face recognition performance. It is a trade-off on how much noise and transformation difference is excluded, and how much intrinsic difference is retained. This eventually leads to the construction of a 3D parameter space that uses the three subspace dimensions as axes. The conventional subspace methods are all limited to local areas of the parameter space. Using this framework, much better recognition can be achieved by searching through the parameter space. However, the strategy to find the optimal parameters was not provided in that work. One possible way is to find the parameters based on experimental evaluation. But it is time consuming and requires to redesign the parameters for different applications.

Because of the complexity of face recognition problem, it is difficult to pursue a single optimal classifier to meet all the requirements. The methods described above all have distinctive shortcomings. Therefore, instead of developing a single optimal classifier, we propose an ensemble learning framework based on random sampling on all the three key components of a classifier: feature space, training samples, and parameter space. The complex face data distribution is learned through multiple subspaces. In our preliminary study (Wang and Tang, 2004), random subspace and bagging were applied to Fisherface and Null Space LDA (N-LDA) and achieved promising performance. In this paper, we conduct a more thorough study, and further extend random sampling to parameter space to complete our framework.

Random subspace (Kam Ho, 1998, 1999) and bagging (Breiman, 1996) are two popular random sampling techniques to enforce weak classifiers. In the random subspace method, a set of low dimensional subspaces is generated by randomly sampling from the original high dimensional feature vector and multiple classifiers constructed in random subspaces are combined in the final decision. In bagging, random bootstrap replicates are generated by sampling the training set. A classifier is constructed from each replicate, and the results of all the classifiers are finally integrated.

Both Fisherface and Null Space LDA (N-LDA) encounter the overfitting problem, but for different reasons. So we will improve them in different ways accordingly. In Fisherface, overfitting happens when the training set is small compared to the high dimensionality of the feature vector. We apply random subspace to reduce the feature vector dimension in order to decrease the discrepancy. In N-LDA, the null space is small when the training sample number is large. This problem can be alleviated by bagging, since each replicate has a smaller number of training samples. Both Fisherface and N-LDA discard some discriminative information. However, the two kinds of classifiers are also complementary, since they are computed in two orthogonal subspaces. Using a fusion rule we combine the two groups of classifiers together to form a more powerful and stable classifier that covers most of the face feature space thus preserves most of the discriminative information. To further boost the performance of the classifier, we also perform random sampling on the parameters of the classification system. This helps us to get around the difficult task of finding the optimal pa-

rameters. Multiple LDA classifiers are constructed by randomly selecting the subspace dimension parameters and combined using a fusion rule. Experiments show that the fusion result is close to the classifier with optimal parameters and it avoids the performance drop because of incorrectly setting the parameters.

We also apply this random sampling approach to the integration of multiple features. Zhao et al. (2003) pointed out that both holistic feature and local features are crucial for face recognition, and have different contributions. Three typical kinds of features, shape, texture, and local Gabor wavelet responses are selected. They undergo a scale normalization and decorrelation by PCA to form a combined long feature vector. Although combining multiple features may worsen the small sample size problem of a traditional LDA, it can be easily resolved under our random sampling framework. The final random sampling face recognition system integrating shape, texture, and Gabor responses achieves 99.83% recognition accuracy for the XM2VTS database.

2. LDA Based Face Recognition

In this section, we briefly review the two conventional LDA face recognition approaches, Fisherface and N-LDA. For appearance-based face recognition, a 2D face image is viewed as a vector with length N in the high dimensional image space. The training set contains M samples $\{\vec{x}_i\}_{i=1}^{M}$ belonging to L individual classes $\{X_j\}_{j=1}^{L}$. LDA tries to find a set of projecting vectors W best discriminating different classes. According to the Fisher criteria, it can be achieved by maximizing the ratio of determinant of the between-class scatter matrix S_b and the determinant of the within-class scatter matrix S_w,

$$W = \arg\max \left| \frac{W^T S_b W}{W^T S_w W} \right|. \tag{1}$$

S_b and S_w are defined as,

$$S_w = \sum_{i=1}^{L} \sum_{\vec{x}_k \in X_i} (\vec{x}_k - \vec{m}_i)(\vec{x}_k - \vec{m}_i)^T, \tag{2}$$

$$S_b = \sum_{i=1}^{L} n_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T, \tag{3}$$

where $\vec{m}_i$ is the mean face for class X_i with n_i samples. W can be computed from the eigenvectors of $S_w^{-1} S_b$ (Fukunnaga, 1991). The rank of S_w is at most M-L. But in face recognition, usually there are only a few samples for each class, and M-L is far smaller than the face vector length N. So S_w may become singular and it is difficult to compute S_w^{-1}.

In the Fisherface method (Belhumeur, 1997), the face data is first projected to a PCA subspace spanned by M-L largest eigenfaces. LDA is then performed in the M-L dimensional subspace, such that S_w is nonsingular. But in many cases, M-L dimensionality is still too high for the training set. When the training set is small, S_w is not well estimated. A slight disturbance of noise on the training set will greatly change the inverse of S_w. So the LDA classifier is often biased and unstable.

Different subspace dimensions are selected in other studies. In Monn and Phillips (1998), the dimension of PCA subspace was chosen as 40% of the total number of eigenfaces. In Swets and Weng (1996), the selected eigenfaces contains 95% of the total energy. They all remove eigenfaces with small eigenvalues. However, eigenvalue is not an indicator of the feature discriminability. In fact, the PCA subspace dimension depends on the training set. When the training set is small, some discriminative information has to be discarded in order to construct a stable LDA classifier.

Chen et al. (2000) suggested that the null space of S_w, in which $W^T S_w W = 0$, also contains much discriminative information. It is possible to find some projection vectors W satisfying $W^T S_w W = 0$ and $W^T S_b W \neq 0$, thus the Fisher criteria in Eq. (1) definitely reaches its maximum value. A LDA approach in the null space of S_w was proposed. First, the null space of S_w is computed as,

$$V^T S_w V = 0. \tag{4}$$

The between-class scatter matrix is projected to the null space of S_w,

$$\tilde{S}_b = V^T S_b V. \tag{5}$$

The LDA projection vectors are defined as $W = V\Phi$, where Φ contains the eigenvectors of $\tilde{S}_b$ with the largest eigenvalues.

N-LDA may also overfit the training set. The rank of S_w, $r(S_w)$ is bounded by $\min(M - L, N)$. Because of the existence of noise, $r(S_w)$ is almost equal

to this bound. The dimension of the null space is $\max(0, N - M + L)$. As shown by experiments in Chen et al., (2000), when the training sample number is large, the null space of S_w becomes small, thus much discriminative information outside this null space will be lost. An extreme case is that the training set is so large that we have $M-L = N$. Then no information can be obtained in this space, since the dimension of the null space is zero.

3. Random Sampling Based LDA for Face Recognition

The above LDA approaches have two common problems: the constructed classifier is unstable and much discriminative information is discarded. In this section, we use random sampling to improve LDA based face recognition. We construct many weak classifiers and combine them into a powerful decision rule. Although Fisherface and N-LDA share the same kind of problems, they are due to different reasons. According to the cause of the problem, we design different random sampling algorithms to improve the two LDA methods. We then combine two improved methods in a multi-classifier structure.

3.1. Random Sampling in Feature Space

Although the dimension of image space is very high, only part of the full space contains discriminant information. This subspace is spanned by all the eigenvectors of the ensemble covariance matrix with nonzero eigenvalues. For the covariance matrix computed from M training samples, there are at most $M-1$ eigenvectors with nonzero eigenvalues. On the remaining eigenvectors with zero eigenvalues, all the training samples have zero projections and no discriminative information can be obtained. Therefore we first project the high dimensional image data to the $M-1$ dimension PCA subspace before random sampling.

In Fisherface, overfitting happens when the training set is relatively small compared to the high dimensionality of the feature vector. In order to construct a stable LDA classifier, we sample a small subset of features to reduce discrepancy between the training set size and the feature vector length. Using such a random sampling method, we construct a multiple number of stable LDA classifiers. We then combine these classifiers to construct a more powerful classifier that covers the entire feature space without losing discriminant information. The proposed random subspace LDA algorithm contains the following steps:

At the training stage,

(1) Apply PCA to the face training set. All the eigenfaces with zero eigenvalues are removed, and $M-1$ eigenfaces $U_t = \{u_1, \ldots, u_{M-1}\}$ are retained as candidates to construct random subspaces.
(2) Generate K random subspaces $\{S_i\}_{i=1}^{K}$. Each random subspace S_i is spanned by $N_0 + N_1$ dimensions. The first N_0 dimensions are fixed as the N_0 largest eigenfaces in U_t. The remaining N_1 dimensions are randomly selected from the other $M - 1 - N_0$ eigenfaces in U_t.
(3) K LDA classifiers $\{C_i(x)\}$ are constructed from the K random subspaces.

At the recognition stage,

(1) The input face data is projected to K random subspaces and fed to K LDA classifiers in parallel.
(2) The outputs of the K LDA classifiers are combined using a fusion scheme to make the final decision.

This algorithm has several novelty features. First, this is the first time that random subspace is applied to face recognition. Second, unlike the traditional random subspace method that samples the original feature vector directly, our algorithm samples in the PCA subspace. We first remove all the eigenfaces with zero eigenvalues, because all the training samples have zero projections on these vectors. The dimension of feature space is first greatly reduced without loss on discriminative information. After PCA, the features on different eigenfaces are uncorrelated, thus are more independent. Better accuracy can be achieved if different random subspaces are more independent from each other (Kuncheva et al., 2001).

Third, our random subspace is not completely random. The random subspace is composed of two parts. The first N_0 dimensions are fixed as the N_0 largest eigenfaces, and the remaining N_1 dimensions are randomly selected from $\{u_{M-N_0-1}, \ldots, u_{M-1}\}$. The N_0 largest eigenfaces encode much face structural information. If they are not included in the random subspace, the accuracy of LDA classifiers may be too low. Although many multiple classifier systems (Kittler and Roli) have been proposed to enforce weak classifiers, the fusion

method will be more complicated if each individual LDA classifier is poor. In our approach, LDA classifier in each random subspace has satisfactory accuracy. The N_1 random diensions cover most of the remaining small eigenfaces. So the ensemble classifiers also have a certain degree of error diversity. Good performance can be achieved using very simple fusion rules such as majority voting.

3.2. *Random Sampling on Training Samples*

Contrary to Fisherface, in N-LDA, the overfitting problem happens when the training sample number is large, since the null space is too small to contain enough discriminative information. This problem can be alleviated by bagging. In bagging, random bootstrap replicates are generated by sampling the training set, so each replicate has a smaller number of training samples. Based on this strategy, we propose the following algorithm:

(1) Apply PCA to the face training set with M samples for L classes. Project all the face data to the $M-1$ eigenfaces $U_t = \{u_1, \ldots, u_{M-1}\}$ with positive eigenvalues.
(2) Generate K bootstrap replicates $\{T_i\}_{i=1}^{K}$. Each replicate contains the training samples of L_1 individuals randomly selected from the L classes.
(3) Construct a N-LDA classifier from each replicate and combine the multiple classifiers using a fusion rule.

Our algorithm randomly selects the individual classes, but does not randomly sample data within each class. This is because in face recognition usually there are a large number of people to be classified but there are very few samples in each class. For example, in our experiment, there are 295 people in the gallery and each personal has only two samples for training. Because human faces share similar intrapersonal variations, the N-LDA constructed from the replicate T_i also can distinguish persons outside T_i, although may not be optimal. The K classifiers can cover all the L classes in the training set.

3.3. *Integrating Random Subspace and Bagging for LDA Based Face Recognition*

While Fisherface is computed from the principal subspace of S_w, in which $W^T S_w W \neq 0$, N-LDA is computed from its orthogonal subspace in which $W^T S_w W = 0$. Both of them discard some discriminative information. Fortunately, the information retained by the two classifiers complements each other. So we combine the two sets of complementary multiple LDA classifiers generated by random sampling to construct the final classifier as illustrated in Fig. 1.

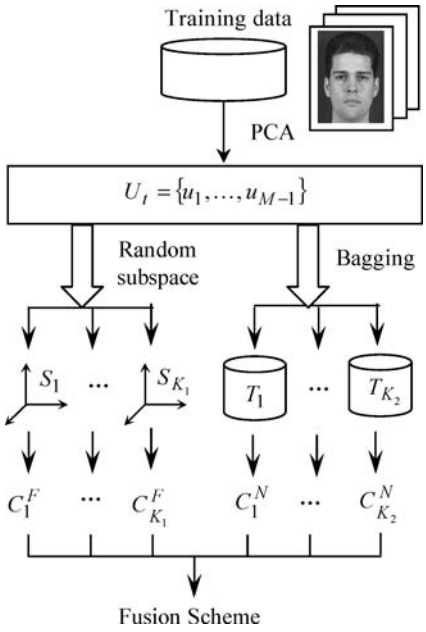

Figure 1. Integrate multiple Fisherface and N-LDA classifiers generated by random sampling. C_i^F is the LDA classifier constructed from the random subspace S_i and C_i^N is the N-LDA classifier constructed from the bagging replicate T_i.

Many methods on combining multiple classifiers have been proposed (Kittler and Roli; Xu et al., 1992; Ross and Jain, 2003; Hong and Jain, 1998; Kegelmeyer and Bowyer, 1997; Kuncheva, 2002; Yacoub et al., 1999). In this paper, we use two simple fusion rules to combine LDA classifiers: majority voting and sum rule. More complex combination algorithms may further improve the system performance.

3.3.1. Majority Voting Each LDA 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,

$$T_k(x \in X_i) = \begin{cases} 1, & C_k(x) = i \\ 0, & otherwise \end{cases}. \tag{6}$$

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

$$\beta(x) = \arg\max_{X_i} \sum_{k=1}^{K} T_k\,(x \in X_i). \tag{7}$$

3.3.2. Sum Rule We assume that $P(X_i|C_k\,(x))$ is the probability that x belongs to X_i under the measure of LDA classifier $C_k\,(x)$. According to the sum rule, the class for the final decision is chosen as,

$$\beta(x) = \arg\max_{X_i} \sum_{k=1}^{K} P\,(X_i|C_k\,(x)) \tag{8}$$

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

$$w_k^i = W_k^T m_i \tag{9}$$

$$w_k^x = W_k^T x \tag{10}$$

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

$$\hat{P}(X_i|C_k\,(x)) = \left(1 + \frac{\left(w_k^x\right)^T \left(w_k^i\right)}{\left\|w_k^x\right\| \cdot \left\|w_k^i\right\|}\right)/2, \tag{11}$$

which has been mapped to [0,1].

3.4. Random Sampling of Parameters

In this section, we consider the problem of subspace dimension parameter for LDA. In Wang and Tang (2004), we proposed a unified framework for subspace analysis. Here we give a brief description to help readers understand this paper, and details can be found in Wang and Tang (2004). In this framework, PCA and Bayesian subspace analysis can be viewed as intermediate steps of LDA, and LDA is performed in three steps:

1. Project face vectors to PCA subspace and adjust the PCA dimension (*dp*) to reduce most noise.
2. Apply Bayesian analysis in the reduced PCA subspace and adjust the dimension (*di*) of intrapersonal subspace to reduce the transformation difference.
3. Project all the face class centers onto the *di* intrapersonal eigenvectors, and then normalize the projections by intrapersonal eigenvalues to compute the whitened class centers. Apply PCA on the whitened class centers to compute a discriminant feature vector of dimensional *dl*. The face class is recognized using the *dl* discriminant features.

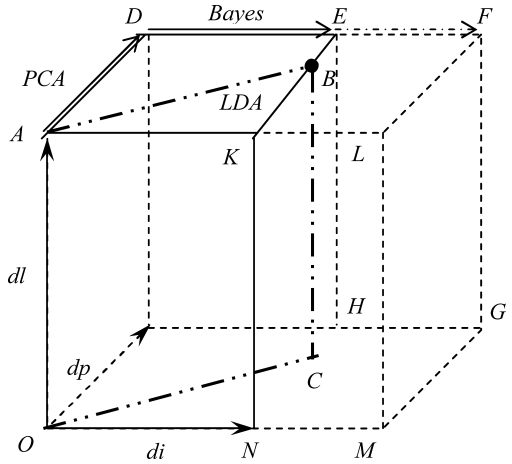

Figure 2. 3D parameter space. *dp*, *di* and *dl* are the dimensionality of PCA subspace, intrapersonal subspace, and LDA subspace.

We could improve each step of subspace analysis by choosing the optimal subspace dimensions in 3D parameter space in Fig. 2. It is a trade-off on how much noise and transformation difference is excluded, and how much intrinsic difference is retained. In conventional subspace methods, the parameters are fixed or constrained to some local regions in Fig. 2. If the selected parameters are not appropriate for a particular data set, the recognition performance may be very poor. We expect to achieve better performance since now the parameters can be freely selected under this framework. However, in Wang and Tang (2004), optimal parameters are only empirically searched by experiments. It is time consuming and requires redesigning the parameters for different applications.

Following the random sampling framework, we construct multiple classifiers performing LDA in three steps. In each step, the subspace dimension is randomly selected. The multiple classifiers are integrated using majority voting. Some individual classifiers with inferior parameters will not seriously deteriorate the system performance because of the ensemble learning. Experiments show that the performance of the multiple classifiers system is close to the classifier with optimal parameters. So our system is more stable and we do not have to pursue the optimal parameters for a particular data set. It can be easily implemented in practice.

3.5. *Theoretical Analysis on the Random Sampling Framework*

A theoretical explanation on how bagging predictors work was given in Breiman (1996). In a similar way, we generalize this discussion to our random sampling framework including random subspace, bagging, and random sampling on parameters. We analyze how different random sampling strategies work in subspace face recognition.

Let x be a sample, and y be the corresponding class label or a score vector indicating the probabilities of x belonging to different classes, $\beta(x, \Gamma)$, where $\Gamma = (T, S, \Omega)$, is a single predictor (classifier) learnt from sample set T, in the feature space S, with parameter setting Ω. If majority voting is used as the fusion rule, $\beta(x, \Gamma)$ is the class label. If sum rule is used, $\beta(x, \Gamma)$ is a score vector with $(0 \ldots 0 \underset{i}{1} 0 \ldots 0)$ indicating that x belongs to class i. Here we assume that $\beta(x, \Gamma)$ is a score vector. A set of different configurations $\{\Gamma_k^{(B)} = (T_k^{(B)}, S_k^{(B)}, \Omega_k)\}$ are generated by random sampling replicate $T_k^{(B)}$ from the whole training set $T^{(A)}$, random subspace $S_k^{(B)}$ from the whole feature space $S^{(A)}$, parameter Ω_k from the parameter space. Aggregate individual predicators as

$$\beta_B(x) = \frac{1}{K}\sum_{k=1}^{K}\beta\left(x, \Gamma_k^{(B)}\right) \approx E_{\Gamma^{(B)}}\left(\beta\left(x, \Gamma^{(B)}\right)\right), \tag{12}$$

where $E_{\Gamma^{(B)}}(\beta(x, \Gamma^{(B)}))$ is the expectation of $\beta(x, \Gamma^{(B)})$ over $\Gamma^{(B)}$. Different configurations $\{\Gamma_k^B\}$ are independently sampled based on an identical distribution. Take x to be a fixed input sample and y the desired output value, and we have

$$E_{\Gamma^{(B)}}\left(y - \beta\left(x, \Gamma^{(B)}\right)\right)^2 = y^2 - 2yE_{\Gamma^{(B)}}\left(\beta\left(x, \Gamma^{(B)}\right)\right) + E_{\Gamma^{(B)}}\left(\beta^2\left(x, \Gamma^{(B)}\right)\right). \tag{13}$$

Since

$$E_{\Gamma^{(B)}}\left(\beta^2\left(x, \Gamma^{(B)}\right)\right) \geq \left[E_{\Gamma^{(B)}}\left(\beta\left(x, \Gamma^{(B)}\right)\right)\right]^2 = [\beta_B(x)]^2, \tag{14}$$

Eq. (13) can derive

$$E_{\Gamma^{(B)}}\left(y - \beta\left(x, \Gamma^{(B)}\right)\right)^2 \geq (y - \beta_B(x))^2. \tag{15}$$

Thus we can get some observations and insights on how to improve subspace face recognition. The mean-squared error of the combined classifier is smaller than the mean-squared error of $\beta\left(x, \Gamma^{(B)}\right)$ under random sampled configuration $\Gamma^{(B)}$, averaged over $\Gamma^{(B)}$. How much improvement we can get depends on the difference between the two sides in (14). If $\beta\left(x, \Gamma^{(B)}\right)$ has large error diversity over $\Gamma^{(B)}$, significant improvement can be obtained by aggregating these classifiers. In Wang and Tang (2004), we have shown that the recognition accuracy has great variation choosing different subspace dimensionalities as parameters. When we do LDA in different random subspaces, we are actually extracting discriminative information in different portions of the face space. The subspace obtained by N-LDA on a training subset is optimal to recognize the people in that replicate, not necessarily optimal to the people in other replicates. So the classifiers obtained under our random sampling framework have large error diversity. Furthermore, because of the fact that LDA is sensitive to noise, a slight change on the configuration $\Gamma^{(B)}$ may lead to different classifiers. Classifier aggregation may overcome the unstability problem of LDA to some extent.

However, it does not mean that we can arbitrarily design the random sampling strategy and expect it to work well in all the cases. Actually, we should compare the performance of $\beta_B(x)$ with that of $\beta\left(x, \Gamma^{(A)}\right)$ instead of $\beta\left(x, \Gamma^{(B)}\right)$. $\Gamma^{(A)} = \left(T^{(A)}, S^{(A)}, \Omega\right)$ includes the whole training set and the whole feature space. If $\beta\left(x, \Gamma^{(B)}\right)$ is much worse than $\beta\left(x, \Gamma^{(A)}\right)$ and the error diversity is not large enough, the random sampling and aggregation strategy may even worsen the recognition accuracy. In our framework, $\Gamma^{(B)}$ and $\Gamma^{(A)}$ share the same parameter space. Without knowledge on the optimal parameters, random sampling on parameters can improve the overall system performance. In random subspace LDA and bagging N-LDA, the individual classifiers generated by random sampling are not much worse or even better than the original classifier using the whole training set in the whole feature space. This is because we correctly design the random sampling strategy according to different problems in LDA and N-LDA. Actually they are less sensitive to noise compared with the original classifier, although they have lost some features or training samples. The problem of LDA is that the training set is small relative to the high dimensional feature vector. Using a smaller feature vector by random sampling the feature space can improve the performance of $\beta\left(x, \Gamma^{(B)}\right)$. In N-LDA, when the training set is very large, the null space is too small to contain enough discriminative information. Using only part of the training data, $\beta(x, \Gamma^{(B)})$ might achieve even better performance than $\beta\left(x, \Gamma^{(A)}\right)$ using the entire

training set. For comparison, it may not be a good strategy to apply bagging to LDA or apply random subspace to null space LDA. Although classifiers generated by random sampling still have large error diversity, they are much worse than the original classifier. In bagging, each replicate has a smaller number of training samples, and thus it makes the small sample problem even worse. Random subspace makes the null space of the within-class scatter matrix smaller. The above analysis will be further explained in our experiments.

3.6. *Discussion on the Random Sampling Framework*

Our random sampling framework is proposed to overcome the overfitting problem and the difficulty of selecting optimal parameters in subspace face recognition. It can be further improved in several aspects. First, the current framework assumes that the face intrapersonal variation has a Gaussian distribution. It works well on the data set without very large intrapersonal variations as shown in the experiments. However, under large pose, illumination changes or occlusion, the manifold of the face intrapersonal variation may be too complicated to be modeled as a single Gaussian distribution. In this case, all the subspace methods based on a single Gaussian model may fail. One possible solution is to extend the current framework to Gaussian mixture models. More detailed discussion can be found in Wang and Tang (2005).

Another problem with the current approach is that we randomly select the classifiers without much consideration on the error diversity. An extra step of classifier selection, for example, using a validation set to measure the classifier error diversity and clustering classifiers based on error diversity, may further improve the system performance (Roli et al., 2001). Finally, the random sampling framework achieves higher and more stable recognition accuracy at higher computational cost. This may become a problem in some online application requiring high speed.

3.7. *Integration of Multiple Features*

Zhao et al. (2003) pointed out that both face holistic features and local features are critical for recognition and have different contributions. We apply this random sampling LDA approach to the integration of multiple features including shape, texture, and Gabor responses.

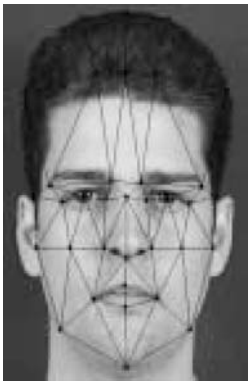

Figure 3. Face graph model.

A face graph containing 35 fiducial points is designed as shown in Fig. 3. Using the method in Active Shape Model (Lanitis, 1997), we separate the face image into shape and texture. The shape vector $\vec{V}_s$ is formed by concatenating the coordinates of the 35 fiducial points after alignment. Warping the face image onto a mean face shape, the texture vector $\vec{V}_t$ is obtained by sampling intensity on the shape-normalized image. As described in Elastic Bunch Graph Matching (Wiskott et al., 1997), a set of Gabor kernels in five scales and eight orientations are convolved with the local patch around each fiducial point. The Gabor feature vector $\vec{V}_g$ combines 35×40 magnitudes of Gabor responses to represent the face local texture. The multi-feature multi-calssifier face recognition algorithm is then designed as following:

- Apply PCA to the three feature vectors respectively to compute the eigenvectors U_s, U_t, U_g and eigenvalues λ_i^s, λ_i^t, λ_i^g. All the eigenvectors with zero eigenvalues are removed.
- For each face image, project each kind of feature to the eigenvectors and normalize them by the sum of eigenvalues, such that they are in the same scale.

$$\vec{w}_j = U_j^T \vec{V}_j / \sqrt{\sum \lambda_i^j}, \ (j = s, \ t, \ g). \qquad (16)$$

- Combine $\vec{w}_t$, $\vec{w}_s$, $\vec{w}_g$ into a large feature vector.
- Apply the random sampling algorithm to the combined feature vector to generate multiple LDA classifiers.

While most other multi-feature integration systems are based on match score level or decision level by designing one classifier for each kind of feature

(Ross and Jain, 2003; Hong and Jain, 1998), our integration approach starts from the feature level. Integration at feature level conveys the richest information, but it is more difficult because of two reasons. First, different kinds of features are incompatible in scale. Second, the new combined feature vector has a higher dimensionality and it will make the small sample size problem even worse. Our approach overcomes both problems. Different features are scaled by PCA normalization and the small sample size problem is resolved by random sampling.

4. Experiments

We conduct experiments on the XM2VTS face database (Messer et al., 1999). There are 295 people, and each person has four frontal face images taken in four different sessions. Some examples are shown in Fig. 4. In our experiments, two face images of each face class are selected for training and reference, and the remaining two for testing. We adopt the recognition test protocol used in FERET (Phillips et al., 1998). All the face classes in the reference set are ranked. We measure the percentage of the "correct answer in top 1 match."

4.1. Random Subspace LDA

We first compare random subspace LDA with the conventional Fisherface approach using the holistic feature. In preprocessing, the face image is normalized by translation, rotation, and scaling, such that the centers of two eyes are in fixed positions. A 46 by 81 mask removes most of the background. So the image space dimensionality is $36 \times 81 = 3726$. Histogram equalization is applied as photometric normalization.

Figure 4. Examples of face images in XM2VTS database taken in four different sessions.

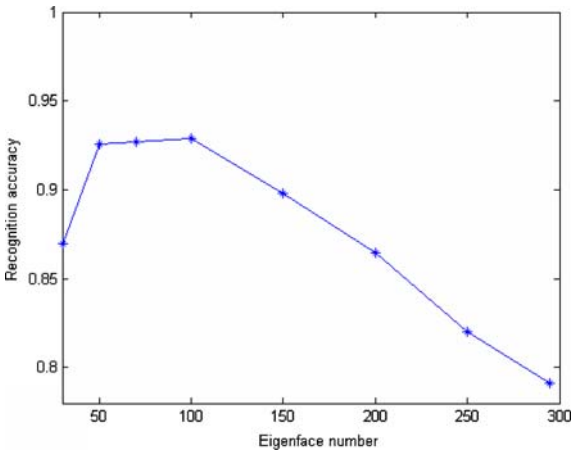

Figure 5. Recognition accuracy of Fisherface classifier using different number of eigenfaces in the reduced PCA subspace.

Figure 5 reports the accuracy of a single LDA classifier constructed from the PCA subspace with different number of eigenfaces. Since there are 590 face images of 295 classes in the training set, there are 589 eigenfaces with non-zero eigenvalues. According to the Fisherface (Wang and Tang, 2004), the PCA subspace dimension should be $M-L = 295$. However, the result shows that the accuracy is only 79% using a single Fisherface classifier constructed from 295 eigenfaces, because this dimension is too high for the training set. We observe that LDA classifier has the best accuracy 92.88% when the PCA subspace dimension is set at 100. So for this data set 100 seems to be a suitable dimension to construct a stable LDA classifier. In the following experiments, we choose 100 as the dimension of random subspaces to construct the multiple LDA classifiers.

First, we randomly select 100 eigenfaces from 589 eigenfaces with nonzero eigenvalues. The result of combining 20 LDA classifiers using majority voting is shown in Fig. 6. With random sampling, the accuracy of each individual LDA classifier is low, between 50% and 70%. Using majority voting, the weak classifiers are greatly enforced, and 87% accuracy is achieved. This shows that LDA classifiers constructed from different random subspaces are complementary of each other. In Fig. 7, as we increase the classifier number K, the accuracy of the combined classifier improves, and

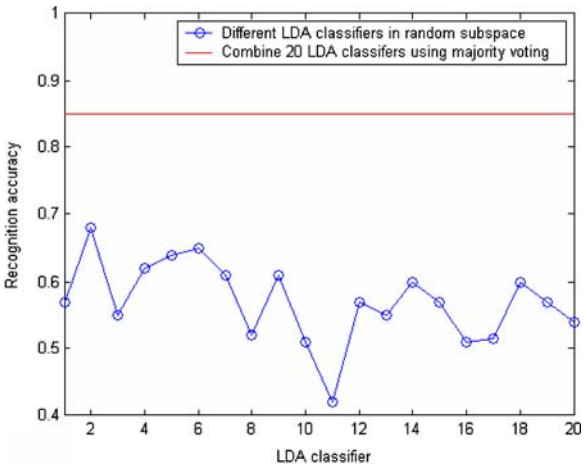

Figure 6. Recognition accuracy of combing 20 LDA classifiers constructed from random subspaces using majority voting. Each random subspace randomly selects 100 eigenfaces from 589 eigenfaces with non-zero eigenvalues.

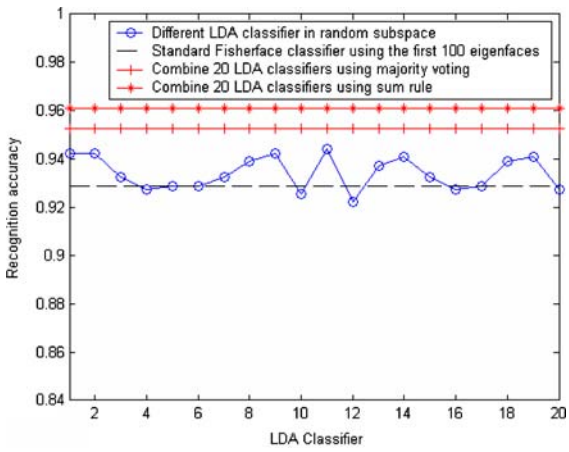

Figure 8. Recognition accuracy of combing 20 LDA classifiers constructed from random subspaces using majority voting and the sum rule. For each 100 dimensional random subspace, the first 50 dimensions are fixed as the 50 largest eigenfaces, and another 50 dimensions are randomly selected from the remaining 539 eigenfaces with non-zero eigenvalues.

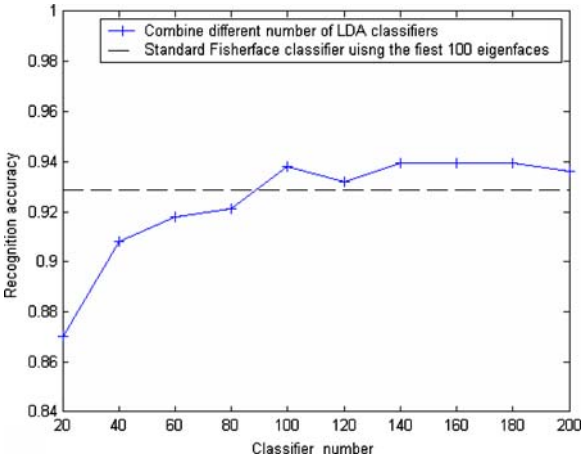

Figure 7. Accuracy of combining different number of LDA classifiers constructed from random subspaces using majority voting. Each random subspace randomly selects 100 eigenfaces from 589 eigenfaces with non-zero eigenvalues.

even becomes better than the highest accuracy in Fig. 5. Although increasing classifier number and using more complex combining rules may further improve the performance, it will increase the system burden.

A better approach to improve the accuracy of the combined classifier is to increase the performance of each individual weak classifier. To improve the accuracy of each individual LDA classifier, as illustrated in Section 3.1, in each random subspace, we fix the first 50 dimensions as the 50 largest eigenfaces, and randomly select another 50 dimensions from the remaining 539 eigenfaces. As shown in Fig. 8, individual LDA classifiers are improved significantly. They are similar to the LDA classifier based on the first 100 eigenfaces. This shows that $\{u_{51}, \ldots, u_{100}\}$ are not necessarily more discriminative than those smaller eignefaces. These classifiers are also complementary of each other, hence much better accuracy is achieved when they are combined. In Table 1, we run the random sampling algorithm 10 times on the same training set and testing set, and then compute the accuracy means and

Table 1. Recognition accuracy of combining LDA classifiers using different number (K) of random subspaces (sum rule). In each random subspace, the first 50 dimensions are fixed as the 50 largest eignfaces, and another 50 dimensions are randomly selected from the remaining 593 eigenfaces with positive eigenvalues. We run ten times on the same training set and testing set, and record the accuracy means and variances.

K	5	10	15	20	25	30
Mean	0.954	0.958	0.959	0.961	0.961	0.962
Variance	0.0133	0.0127	0.0094	0.0101	0.0068	0.0049

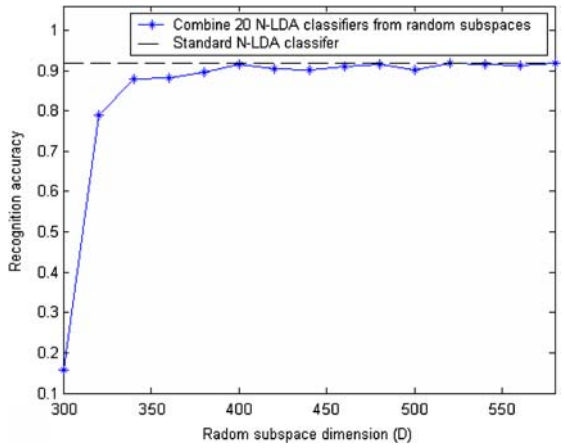

Figure 9. Recognition accuracy of combining 20 N-LDA classifiers from random subspaces.

variances. Using more random subspaces, the accuracy is higher and more stable.

We also apply random subspace to N-LDA. Similar to the method in Section 3.1, the random subspaces with dimension D ($295 < D < 590$) are generated from PCA subspace and a N-LDA classifier is constructed from each random subspace. As shown in Fig. 9, there is no improvement in recognition performance. When the random subspace dimensionality D is low, the null space dimension (D-295) is small, so the recognition accuracy drops greatly.

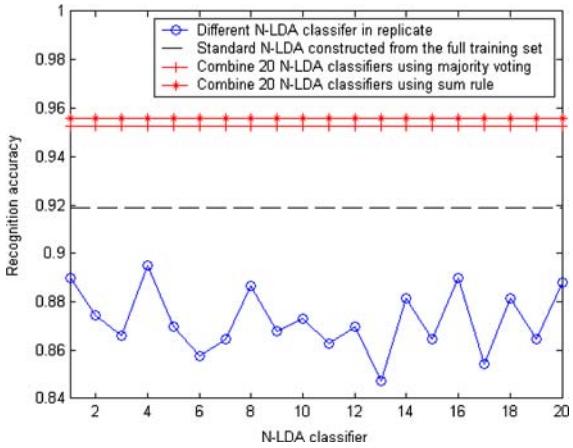

Figure 10. Recognition accuracy of combining 20 N-LDA classifiers constructed from bagging replicates using majority voting and sum rule. Each replicate contains 150 training people.

Table 2. Recognition accuracy of combining N-LDA classifiers using different number (K) of bagging replicates (sum rule). We run ten times on the same training set and testing set, and record the accuracy means and variances.

K	5	10	15	20	25	30
Mean	0.929	0.934	0.942	0.956	0.951	0.961
Variance	0.0120	0.0109	0.097	0.009	0.036	0.027

4.2. *Bagging LDA*

Figure 10 reports the performance of bagging based N-LDA. We generate 20 replicates and each replicate contains 150 people for training. As expected, the individual N-LDA classifier constructed from each replicate is less effective than the original classifier trained on the full training set. However, when the multiple classifiers are combined, the accuracy is significantly improved, and becomes much better than the original N-LDA. Table 2 reports performance of bagging based N-LDA using different number of replicates, but fixing training sample number in each replicate as 300. Similar to results in Table 1, the results are more stable using a relatively large number of replicates.

We also study using bagging to improve Fisherface classifiers. The PCA subspace is spanned by the 100 largest eigenfaces and 20 replicates are generated. The accuracies with the replicate containing different number of people are shown in Fig. 11. As expected, the combined classifier shows no improvement over the original Fisherface classifier.

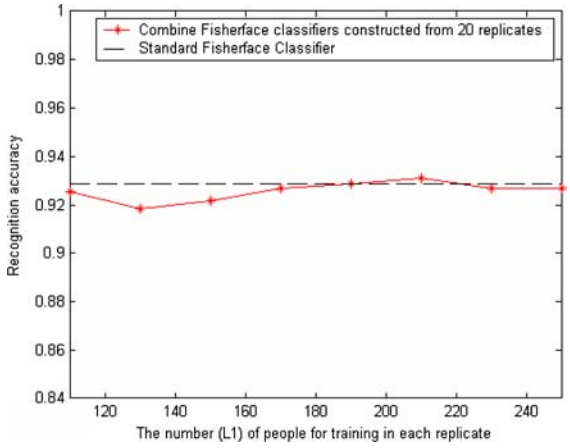

Figure 11. Recognition accuracy of combining 20 Fisherface classifiers constructed from bagging replicates containing different number (L) of people for training. The PCA space is spanned by 100 largest eigenfaces. The combining rule is majority voting.

Table 3. Compare random sampling based LDA with conventional methods. R-LDA (1): random subspace based Fisherface; R-LDA (2): bagging based N-LDA; R-LDA (3): unified subspace analysis by random sampling on parameter space. R-LDA (4): integration of random sampling on feature space, training samples and parameter space.

Feature	Method	Accuracy (%)
Holistic feature	Eigenface	85.59
	Fisherface	92.88
	Bayes	92.71
	R-LDA (1)	96.10
	R-LDA (2)	95.59
	R-LDA (3)	97.12
	R-LDA (4)	98.47
Texture	Euclid distance	85.76
Shape	Euclid distance	49.50
Gabor	EBGM	95.76
Integration of multi-feature	R-LDA (4)	99.83

Integrating the multiple Fisherface classifiers generated by random subspace and N-LDA classifiers generated by bagging, the recognition accuracy can be further improved. We combine 10 Fisherface classifiers constructed from random subspaces and 10 N-LDA classifiers constructed from bagging replicates, and get an even better result as shown in Table 3.

4.3. Unified Subspace Analysis Based on Random Sampling in Parameter Space

In this section, we discuss the experiment on random sampling in the parameter space. We still use the holistic feature as in Section 4.1 and 4.2. Figure 12 plots the recognition accuracy of combining 20 LDA classifiers random sampling in the parameter space. As described in unified subspace analysis (Wang and Tang, 2004), for each classifier we perform LDA by three steps, but randomly select the subspace dimension in each step. Choosing different parameters, the LDA classifiers have great variation. That explains the importance of parameter selection in LDA. With inferior parameters, the accuracy of LDA is even lower than 50%. However, in our random sampling framework, the inferior parameter selection does not affect the system performance much. The result of combining 20 LDA classifiers is even better than the best LDA classifier with good parameters. So the random sampling framework successively boosts the system performance and solves the parameter selection problem.

4.4. Integration of Multiple Features

In Table 3, we report the recognition accuracy of integrating shape, texture, and Gabor features using

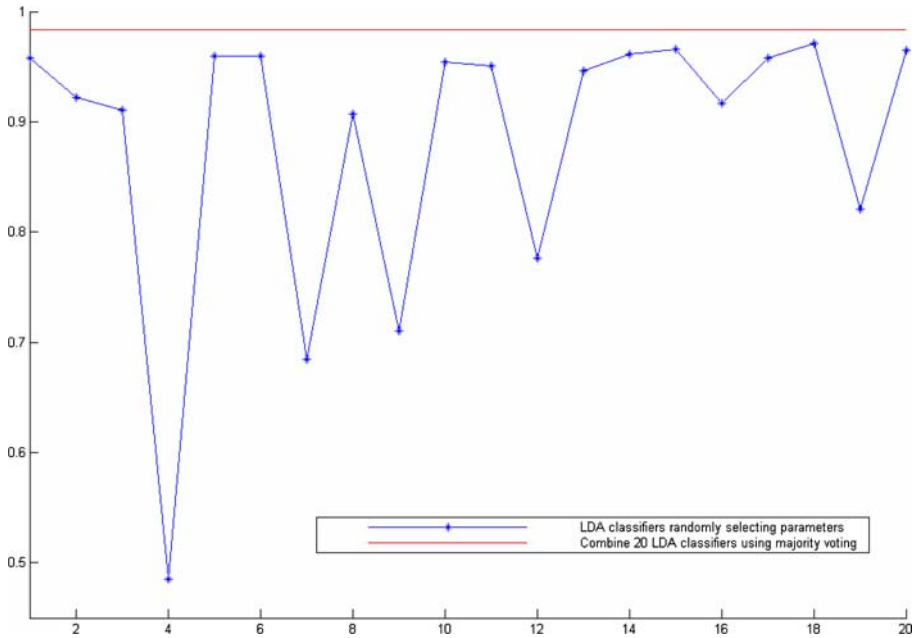

Figure 12. Recognition accuracy of combining 20 LDA classifiers random sampling in the parameter space.

random sampling LDA. Combining 20 classifiers using the sum rule, we achieve 99.83% recognition accuracy. For 590 testing samples, it misclassifies only one! For comparison, we also compute the accuracies of some conventional face recognition approaches in Table 3. Eigenface (Turk and Pentland, 1991), Fisherface (Wang and Tang, 2004), and Bayesian analysis (Moghaddam et al., 2000) are three subspace face recognition approaches based on holistic feature. Elastic Bunch Graph Matching as described in (Wiskott et al., 1997) uses the correlation of Gabor features as similarity measure. Experiments clearly demonstrate the superiority of our new algorithm.

5. Conclusion

Face recognition is a challenging pattern recognition problem. Most previous researches focus on pursuing a single optimal classifier. In this paper, we suggest an alternative approach based on ensemble learning, i.e. using multiple classifiers to solve the complex problem. We develop a robust face recognition system by randomly sampling feature space, training samples, and parameter space. In this paper, our discussion focuses on the LDA method, however, the random sampling framework can also be extended to other subspace methods based on similar consideration. In this study, we use the simplest fusion rules to combine multiple LDA classifiers and achieve notable improvement. Many more complex combination algorithms (Kittler and Roli) have been proposed. They may further improve the performance. Using random subspace, a large set of LDA classifiers can be generated. Instead of combining them directly, it is helpful to select a small set of complementary LDA classifiers with high accuracy for combination. This is a direction for our further study.

Acknowledgements

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region and a joint grant (N_CUHK409/03) from HKSAR RGC and China NSF. The work was done while all the authors are with the Chinese University of Hong Kong.

References

Belhumeur, P.N., Hespanda, J., and Kiregeman, D. 1997. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. on PAMI,* 19(7):711–720.

Breiman, L. 1996. Bagging Predictors. *Machine Learning*, 24(2): 123–140.

Chen, L., Liao, H., Ko, M., Liin, J., and Yu, G. 2000. A New LDA-based Face Recognition System Which can Solve the Samll Sample Size Problem. *Pattern Recognition*, 33(10): 1713–1726.

Fukunnaga, K. 1991. Introduction to Statistical Pattern Recognition. Academic Press, second edition.

Hong, L. and Jain, A.K. 1998. Integrating Faces and Fingerprints for Personal Identification. *IEEE Trans. on PAMI*, 20(12):1295–1307.

Kam Ho, T. 1999. Nearest Neighbour in Random Subspace. *Intelligent Data Analysis*, 3:191–209.

Kam Ho, T. 1998. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. on PAMI*, 20(8):832–844.

Kegelmeyer, W.P. and Bowyer, K. 1997. Combination of Multiple Classifier Using Local Accuracy Estimates. *IEEE Trans. on PAMI*, 19(4):405–410.

Kittler, J. and Roli, F. (Eds):Multiple Classifier Systems.

Kuncheva, L.I. 2002. Switching Between Selection and Fusion in Combining Classifiers: An Experiment. *IEEE Trans. on Systems, Man and Cybernetics, Part B*, 32(2).

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., and Duin, R.P.W. 2001. Is Independence Good for Combining Classifiers? *Proc. of ICPR*, 2:168–171.

Lanitis, A., Taylor, C.J., and Cootes, T.F. 1997. Automatic Interpretation and Coding of Face Images Using Flexible Models. *IEEE Trans. on PAMI*, 19(7):743–756.

Lu, X. and Jain, A.K. 2003. Resampling for Face Recognition. *Proceedings of the 4th International Conference on Audio- and Video-Based Personal Authentication*. Guildford, UK, pp. 869–877.

Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G. 1999. XM2VTSDB: The Extended M2VTS Database. *Proceedings of International Conference on Audio- and Video-Based Person Authentication*, pp. 72–77.

Moghaddam, B., Jebara, T., and Pentland, A. 2000. Bayesian Face Recognition. Pattern Recognition, 33:1771–1782.

Monn, H. and Phillips, P.J. 1998. Analysis of PCA-Based Face Recognition Algorithms. *Empirical Evaluation Techniques in Computer Vision*, Bowyer, K.W. and Phillips, P.J. (eds.), *IEEE Computer Society Press*, Los Alamitos, CA.

Phillips, P.J., Moon, H., Rizvi, S.A., and Rauss, P.J. 1998. The FERET Evaluation. In *Face Recognition: From Theory to Applications*, Wechsler, H., Phillips, P.J., Bruce, V., Soulie, F.F., and Huang, T.S. (eds.), Berlin: Springer-Verlag.

Roli, F., Giacinto, G., and Vernazza, G. 2001. Methods for Designing Multiple Classifier Systems. In *Proceedings of the Second International Workshop on Multiple Classifier Systems*, pp. 78–87.

Ross, A. and Jain, A. 2003. Information Fusion in Biometrics. *Pattern Recognition Letters*, 24:2115–2125.

Swets, D., Weng, J. 1996. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Trans. on PAMI*, 16(8):831–836.

Turk, M. and Pentland, A. 1991. Face recognition using eigenfaces. *Proceedings of IEEE, CVPR*, Hawaii, pp. 586–591.

Wang, X. and Tang, X. 2005. Subspace Analysis Using Random Mixture Models. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*.

Wang, X. and Tang, X. 2004. Dual-Space Linear Discriminant Analysis for Face Recognition. *In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA, pp. 564–569.

Wang, X. and Tang, X. 2004. Random Sampling LDA for Face Recognition. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, pp. 259–265,

Wang, X. and Tang, X. 2004. A Unified Framework for Subspace Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1222–1228.

Wiskott, L., Fellous, J.M., Kruger, N., and von der Malsburg, C. July 1997. Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):775–779.

Xu, L., Krzyzak, A., and Suen, C.Y. 1992. Method of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. *IEEE Trans. on System, Man, and Cybernetics*, 22(3):418–435.

Yacoub, S.B., Abdeljaoud, Y., and Mayoraz, E. 1999. Fusion of Face and Speech Data for Person Identity Verification. *IEEE Transactions on Neural Networks*, 10(5):1065–1074.

Zhao, W., Chellappa, R., Phillips, A.P.J., and Rosenfeld. 2003. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458.