Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 1470–1482
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Selecting discriminant eigenfaces for face recognition

Jie Wang *, K.N. Plataniotis, A.N. Venetsanopoulos

The Edward S. Rogers Sr., Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario, Canada M5A 3G4

Received 16 April 2004; received in revised form 23 July 2004
Available online 11 January 2005

Communicated by H. Wechsler

Abstract

In realistic face recognition applications, such as surveillance photo identification, supervised learning algorithms usually fail when only one training sample per subject is available. The lack of training samples and the considerable image variations due to aging, illumination and pose variations, make recognition a challenging task. This letter proposes a development of the traditional eigenface solution by applying a feature selection process on the extracted eigenfaces. The proposal calls for the establishment of a feature subspace in which the intrasubject variation is minimized while the intersubject variation is maximized. Extensive experimentation following the FERET evaluation protocol suggests that in the scenario considered here, the proposed scheme improves significantly the recognition performance of the eigenface solution and outperforms other state-of-the-art methods.

Keywords: Face recognition; Eigenface selection; Intersubject variation; Intrasubject variation

1. Introduction

Face recognition (FR) is one of the most active research areas in computer vision and pattern recognition with practical applications that include forensic identification, access control and human computer interface. The task of a FR system is to compare an input face image against a database containing a set of face samples with known identity and identify the subject to which the input face belongs. However, a straightforward implementation is difficult since faces exhibit significant variations in appearance due to acquisition, illuminations, pose and aging variations (Adini et al., 1997). In applications such as surveillance photo identification and forensic identification, the subject of interest, photographed in an uncontrolled environment, may significantly differ from possible,

* Corresponding author. Tel.: +1 416 978 2934; fax: +1 416 978 4425.
E-mail address: jwang@dsp.utoronto.ca (J. Wang).

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.11.029

if any, templates stored in the database, mostly due to the lighting and pose variations. In addition, since the timely update of forensic face templates is almost impossible under realistic conditions, appearance deviations due to aging is unavoidable, further complicating the FR tasks. If at the same time, there is only a limited number of training samples available, the characterization of the intrinsic properties of the subject becomes a difficult task. When only one image per subject is available, the problem requires particular attention.

In literature, numerous FR algorithms have been proposed and the state-of-the-art in the area is reported in a series of recent surveys (Chellappa et al., 1995; Zhao et al., 2003). Among the various FR procedures, appearance based solutions, which treat the 2D face image as a vector in the image space, seem to be the most successful (Brunelli and Poggio, 1993). In general, the whole recognition procedure includes two steps, feature extraction and classification. During the feature extraction stage, the original image space is projected to a much lower dimensional feature subspace by using the subspace techniques such as principle component analysis PCA (eigenface) (Turk and Pentland, 1991; Perlibakas, 2004; Yang et al., 2004), independent component analysis (ICA) (Bartlett et al., 2002; Liu and Wechsler, 1999), linear discriminant analysis LDA (fisherface) (Belhumeur et al., 1997; Kim et al., 2003; Yang and Yang, 2003; Lu et al., 2003) and so on. PCA is based on Gaussian models which could separate second order dependencies among two pixels while ICA, a generalization of PCA, could separate high-order moments of the input image. Both PCA and ICA are unsupervised learning techniques which compress the data without considering the class label even if they are available. However, LDA, as a supervised technique, is a class specific solution and searches for the feature basis vectors on which the ratio of the between class and within class scatters is maximized. Upon the extraction of the proper set of features, a classifier such as nearest neighbor classifier, Bayesian classifier (Duda et al., 2000; Jain et al., 2000), neural network (Jain et al., 2000; Er et al., 2002), support vector machine (Burges, 1998), is applied to recognize the face images.

In most of the feature extraction methodologies, the eigenface (PCA) approach and fisherface (LDA) approach are two of the most commonly used subspace techniques. Eigenface, which is based on the Karhunen–Loeve transform, produces an expressive subspace for face representation and recognition while fisherface produces a discriminating subspace. For the purpose of classification, LDA is generally believed to be superior to PCA when enough training samples per subject are available (Belhumeur et al., 1997; Lu et al., 2003). However, when the number of available training samples per subject is small, experimental analysis indicates that PCA outperforms LDA (Martez and Kak, 2001; Beveridge et al., 2001). In particular, when only one training sample per subject is available, the problem considered in this letter, LDA can not be readily applied, since the within class scatter can not be estimated using only one sample per subject.

In this work, we propose a solution based on the eigenface approach. It is well known that in the eigenface approach, the extracted eigenspace maximizes not only the intersubject variation but also the intrasubject variation. This is due to the fact that the eigenface solution works as an unsupervised technique without considering the class label, coupling together both the inter and intra subject variation (Wang and Tang, 2003). However, for classification purposes, intrasubject variation is expected to be small compared to intersubject variation. Therefore the variation from one image to the next is attributed mostly to the subject identity, simplifying the classification problem (Belhumeur et al., 1997). To that end and in order to make the eigenface approach more attractive to classification tasks, we propose to enhance the traditional eigenface solution by applying a feature selection process on the extracted eigenfaces. The objective is to select those eigenfaces that form a subspace in which intersubject variation is maximized and intrasubject variation is minimized. Thus, the obtained feature subspace becomes more discriminant for classification tasks making the recognition performance significantly improved.

In order to demonstrate the validity of the proposed solution, under the above mentioned scenario, the well known FERET database is used

in our experimentation (Phillips et al., 2000). Extensive simulation studies on the FERET database indicate that the proposed feature selection scheme improves significantly the recognition performance when large intrasubject variation exists in the probe images, the application scenario most often encountered in practical FR tasks.

The rest of the paper is organized as follows: Section 2 formulates the problem and introduces the proposed framework. In Section 3, the eigenface method is briefly reviewed for completeness. The selection criterion and the procedure developed are introduced and analyzed in Section 4. Motivations and design issues are also discussed. Experimental results obtained using the FERET database are given in Section 5. Section 6 summarizes the findings of this work.

2. Problem formulation and system framework

For the application considered here, the face image of the subject of interest is fed to the FR system, which is asked to return the stored examples from the database which match most closely the input, along with the corresponding identities. Using FERET terminology, the problem is stated as follows: Given a set of N images $\{\boldsymbol{x}_i\}_1^N$ along with their identities $l(\boldsymbol{x}_i)$, each of which is represented as a vector of length $L = I_w \times I_h$, i.e., $\boldsymbol{x}_i \in \mathbb{R}^L$, where $(I_w \times I_h)$ is the image size, the task of the FR system is to determine the identity of the subject shown in the input image $\boldsymbol{p}$. Following the FERET naming convention, the input image $\boldsymbol{p}$ will be referred as probe image and images with known identities are named as gallery images. To simulate a realistic operating environment, it will be assumed that there is no overlap between the probe and the gallery. Further to that, it will be assumed that each subject in gallery is represented by a single frontal image.

To determine the identity of a probe face, the probe is compared with each gallery image by calculating the corresponding distance in an "optimal" feature subspace, in which the classes are supposed to be well separated. The gallery images reporting the smallest distances, in the feature space, are selected as candidates for subject identification. The reason for returning more than one candidate instead of the top one match lies in the fact that when the FR task is difficult, namely, face appearance exhibits large variations while only limit training samples are available, top one recognition rate is far from satisfaction for realistic applications. The detailed results will be discussed in the experiment section. The diagrammatic representation of the procedure is given in Fig. 1.

In order to find such an optimal subspace, we propose to employ a feature selection mechanism based on the eigenface approach. We start by collecting a generic training set, from which a set of eigenfaces are extracted. Although image samples

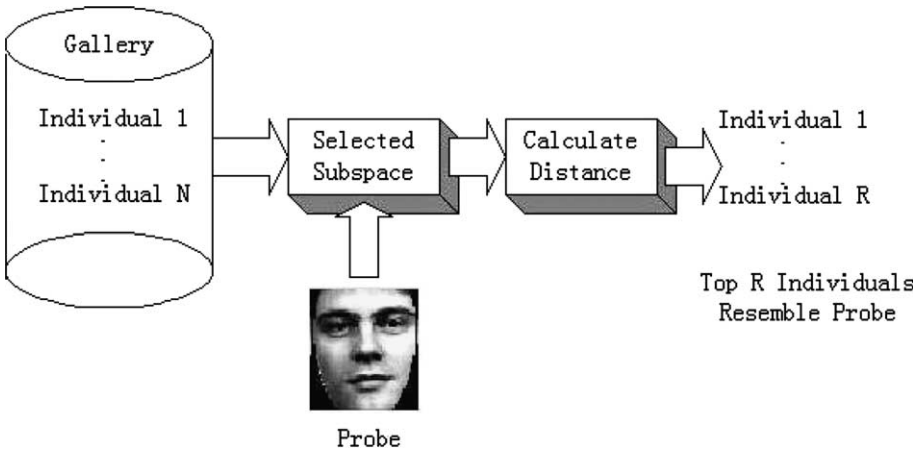

Fig. 1. System framework.

for gallery subjects are limited, there are plenty of face images for other subjects available for training. Therefore, the generic training set can be collected from any available face databases, as long as the subjects included do not overlap with those available in either the gallery or the probe, which matches the conditions in realistic applications. Then the selection procedure is applied on the extracted eigenfaces with the criterion to maximize the intersubject variation and minimize the intrasubject variation as well. The diagram of the selection procedure is illustrated in Fig. 2.

3. Review of eigenface (PCA) method

In the eigenface method, PCA is used to determine the eigenvectors of sample covariance matrix $C = \sum_{k=1}^{N}(\boldsymbol{x}_k - \mu)(\boldsymbol{x}_k - \mu)^{\mathrm{T}}$, where $\mu = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k$ is the mean of all face samples. In the standard eigenface approach, the first m eigenvectors u_i, $i = 1, \ldots, m$, corresponding to the m largest eigenvalues are forming the eigenspace $U = [u_1, \ldots, u_m]$, in which the subsequent recognition is performed.

The eigenface approach is an unsupervised linear technique which provides an optimal, in the mean square error sense, representation of the input in a lower dimensional space (Belhumeur et al., 1997). It produces the most expressive subspace for face representation but not necessarily be the most discriminating one. This is due to the fact that the sample covariance matrix C includes all face difference pairs, those belonging to the same individual and those belonging to different individuals (Wang and Tang, 2003):

$$\begin{aligned} C &= \sum_{k=1}^{N}(\boldsymbol{x}_k - \mu)(\boldsymbol{x}_k - \mu)^{\mathrm{T}} \\ &= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \\ &= \frac{1}{2N}\sum_{l(\boldsymbol{x}_i)=l(\boldsymbol{x}_j)}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \\ &\quad + \frac{1}{2N}\sum_{l(\boldsymbol{x}_i)\neq l(\boldsymbol{x}_j)}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \end{aligned} \tag{1}$$

The eigenspace computed by maximizing the scatter matrix C includes two coupled together factors, intersubject variation and intrasubject variation. The first m eigenfaces corresponding to the largest eigenvalues contain not only large intersubject variation (useful for classification purposes), but also large intrasubject variation. Inclusion of

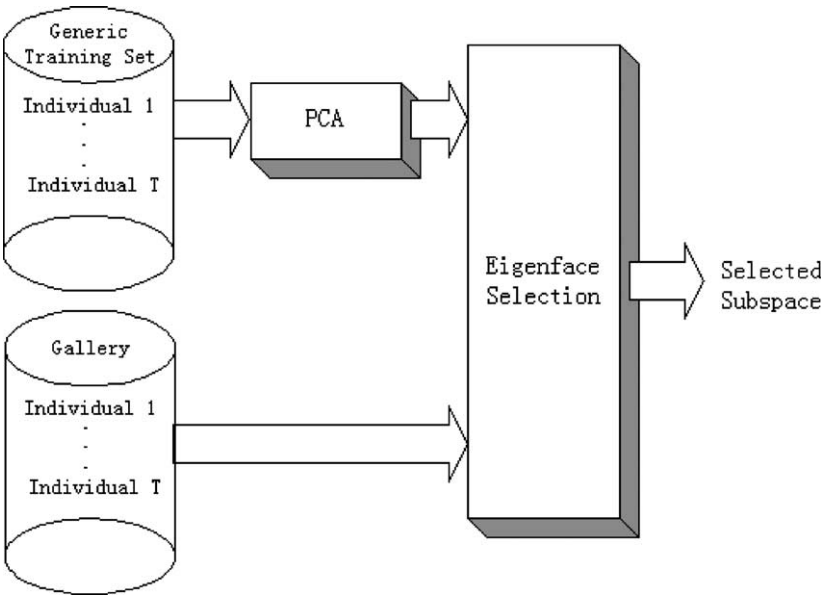

Fig. 2. Training session.

large intrasubject variation, however, is harm to the classification. Therefore, in the eigenspace, resulted from Karhunen–Loeve transform, the face classes can not be guaranteed to be well separated and clustered due to its large intrasubject variation. At the same time, the discarded eigenvectors corresponding to small eigenvalues may carry important discriminant information due to its small intrasubject variation. As a result, some face samples may deviate from the corresponding class centers and be more closer to the centers of other subjects (Etemad and Chellappa, 1997), which will create the problem in the classification stage. On the other hand, if only the eigenfaces corresponding to small eigenvalues are selected, the intrasubject variation will be greatly reduced, however, the discriminating information of intersubject variation will be lost (Belhumeur et al., 1997).

Therefore selecting eigenfaces corresponding to the dominant eigenvalues following the traditional eigenface paradigm or discarding several principle components may not be appropriate for recognition. It is therefore, necessary to develop a method to systematically select the most discriminant eigenfaces from the set of eigenfaces created by the PCA.

4. Select eigenfaces

4.1. Selection criterion

Let $\mathscr{G}$ be the gallery set with G face images (one per subject), $\boldsymbol{g}_i$, $i = 1, 2, \ldots, G$. Let $\mathscr{T}$ be the generic training set of size $S \times L$ containing S subjects, L face images each. $\boldsymbol{t}_{i,j}$ is the jth image of identity i, $j = 1, 2, \ldots, L$; $i = 1, 2, \ldots, S$. PCA is applied on the generic training set. Therefore at most $S \times L - 1$ meaningful eigenvectors with non zero eigenvalues can be obtained by PCA when the number of training samples is less than the dimensionality of the image. Other than selecting from all available eigenfaces, the first M eigenfaces corresponding to the largest eigenvalues are kept to form the eigenface set for selection. Define $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M]$ as the complete eigenface set sorted in descending order of the corresponding eigenvalues, from which the selection is performed. The cardinality of A, namely M, is chosen such that $\sum_{k=1}^{M} \lambda_k / \sum_{k=1}^{S \times L - 1} \lambda_k$ is greater than a threshold, where λ_i is the ith eigenvalue. The reason for exclusion of trailing eigenfaces is due to the fact that eigenfaces corresponding to small eigenvalues are usually unreliable since limited number of samples are used for training. Therefore the problem is reduced to select a subset A_m with cardinality m from the complete eigenface set A which optimizes a selection criterion $J(\cdot)$. In the standard eigenface method, maximization of the selection criterion $J = (\sum_{k=1}^{m} \lambda_k)$, results in $A_m = A_{1:m} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]$.

From a classification point of view, the difference of two face image vectors is expected to be only due to subject identities, so that the selected feature subspace should be the one contains large intersubject variation and small intrasubject variation as well. Therefore, we propose a selection criterion which based on the maximization of the following ratio:

$$J = \frac{\mathrm{Var}_{\mathrm{inter}}(A_m)}{\mathrm{Var}_{\mathrm{intra}}(A_m)} \tag{2}$$

where $\mathrm{Var}_{\mathrm{inter}}(A_m)$ and $\mathrm{Var}_{\mathrm{intra}}(A_m)$ represent intersubject and intrasubject variation in the eigenspace spanned by the eigenfaces in the feature set A_m respectively. Since subjects in the gallery set are to be identified, both the intra and inter variations, namely $\mathrm{Var}_{\mathrm{intra}}(A_m)$ and $\mathrm{Var}_{\mathrm{inter}}(A_m)$, should be preferably estimated using gallery samples. However, in the scenario under consideration here, for each gallery subject, there is only one image is available. Therefore, estimation of intrasubject variation from gallery samples is impossible. Based on the assumption that human faces exhibit similar intraclass variation (Wang and Tang, 2003), the intrasubject variation of the stored gallery subjects are to be estimated from the collected generic training samples, denoted as $\mathrm{Var}_{\mathrm{intra:train}}(A_m)$.

As for intersubject variation, it is expected to characterize the variations for gallery subjects specifically. Therefore, estimating using only gallery images is an appropriate choice, i.e., $\mathrm{Var}_{\mathrm{inter}} = \mathrm{Var}_{\mathrm{inter:gallery}}$. However, due to the limit sample size for each gallery subject, such estimation relies heavily on the stored examples, giving rise to high variance (Duda et al., 2000). On the contrary, if face images of other subjects are included to

reduce the estimation variance, the estimated intersubject variation includes not only the variations among gallery subjects but also those of other subjects, making the selected feature subspace bias the optimal one which aims at the discrimination of the gallery subjects only. Therefore we propose to estimate the intersubject variation by using both generic training ($\mathrm{Var}_{\mathrm{inter:train}}(A_m)$) and gallery samples ($\mathrm{Var}_{\mathrm{inter:gallery}}(A_m)$) with a regularization factor η to balance the bias and variance of the estimation, which is:

$$J = \frac{\eta \mathrm{Var}_{\mathrm{inter:train}}(A_m) + (1-\eta)\mathrm{Var}_{\mathrm{inter:gallery}}(A_m)}{\mathrm{Var}_{\mathrm{intra:train}}(A_m)} \tag{3}$$

If $\eta = 0$, only gallery images are used to determine the inter variation, in which J has zero bias but exhibits large variance. When $\eta = 1$, generic training image set dominates the selection resulting in a biased solution.

4.2. Determining parameters

In the following, we discuss the detailed determination of both inter and intra variation using generic training and gallery samples. From Eq. (1), we know that the eigenfaces are computed by diagonalizing the total scatter C, which is the covariance matrix for all image difference pairs $\{(\boldsymbol{x}_i - \boldsymbol{x}_j)\}$. Let us define I as the intrasubject difference set, i.e., $I{:}\{(\boldsymbol{x}_i - \boldsymbol{x}_j)|l(\boldsymbol{x}_i) = l(\boldsymbol{x}_i)\}$, and E as the intersubject difference set, i.e., $E{:}\{(\boldsymbol{x}_i - \boldsymbol{x}_j)|l(\boldsymbol{x}_i) \neq l(\boldsymbol{x}_j)\}$. Set I and set E are zero-mean, since for each $(\boldsymbol{x}_i - \boldsymbol{x}_j)$, there exists a $(\boldsymbol{x}_j - \boldsymbol{x}_i)$ (Moghaddam et al., 2000). Let C_I and C_E be the corresponding covariance matrices, i.e.,

$$\begin{aligned} C_I &= \frac{1}{N_I}\sum_{l(\boldsymbol{x}_i)=l(\boldsymbol{x}_j)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \\ C_E &= \frac{1}{N_E}\sum_{l(\boldsymbol{x}_i)\neq l(\boldsymbol{x}_j)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \end{aligned} \tag{4}$$

where N_I and N_E are the number of difference pairs in set I and set E. Therefore, the intrasubject and intersubject variation in the direction defined by the basis $\boldsymbol{a}_i$ can be estimated from $\mathrm{Var}_{\mathrm{intra}}(\boldsymbol{a}_i) = \boldsymbol{a}_i^{\mathrm{T}} C_I \boldsymbol{a}_i$, $\mathrm{Var}_{\mathrm{inter}}(\boldsymbol{a}_i) = \boldsymbol{a}_i^{\mathrm{T}} C_E \boldsymbol{a}_i$ respectively. Since the classification is performed by calculating the Euclidean distance in the selected feature subspace, which actually sums the squared difference vector on the eigenbases spanning the feature subspace. Thus, only the variations on the selected eigenbases affect the recognition performance and variations in the selected subspace A_m are therefore reasonably estimated from the summation of variances on the included eigenbases, i.e., $\mathrm{Var}_{\mathrm{intra}} = \mathrm{trace}(A_m^{\mathrm{T}} C_I A_m)$ and $\mathrm{Var}_{\mathrm{inter}} = \mathrm{trace}(A_m^{\mathrm{T}} C_E A_m)$.

It can be proved that, C_I and C_E are actually equivalent to the well known within class scatter S_{w} and between class scatter S_{b} in LDA approaches (Moghaddam et al., 2000; Wang and Tang, 2003). Therefore the proposed selection criterion of Eq. (2) is similar to Fisher criterion. The major differences lie in the two facts (1) $S_{\mathrm{w}}(C_I)$ and $S_{\mathrm{b}}(C_E)$ in the proposed criterion are estimated by using different training sets, generic training set and gallery according to Eq. (3); (2) other than maximizing the ratio of S_{b} and S_{w}, the proposed criterion is to maximize the ratio of their traces.

So, for intersubject variation for generic training samples,

$$\mathrm{Var}_{\mathrm{inter:train}}(A_m) = \sum_{i=1}^{m} \Lambda_{E:\mathrm{train}}(i) \tag{5}$$

$$\Lambda_{E:\mathrm{train}} = \mathrm{diag}\{A_m^{\mathrm{T}} S_{\mathrm{b:train}} A_m\} \tag{6}$$

$$S_{\mathrm{b:train}} = \frac{1}{S}\sum_{k=1}^{S} (\bar{\boldsymbol{m}}_k - \bar{\boldsymbol{m}})(\bar{\boldsymbol{m}}_k - \bar{\boldsymbol{m}})^{\mathrm{T}} \tag{7}$$

where $\bar{\boldsymbol{m}}_k = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{t}_{ki}$ is the mean of face subject k and $\bar{\boldsymbol{m}} = \frac{1}{LS}\sum_{k=1}^{S}\sum_{i=1}^{L} \boldsymbol{t}_{ki}$ is the mean of all generic training samples. Similarly, intrasubject variation for generic training samples can be calculated as follows:

$$\mathrm{Var}_{\mathrm{intra:train}}(A_m) = \sum_{i=1}^{m} \Lambda_{I:\mathrm{train}}(i) \tag{8}$$

$$\Lambda_{I:\mathrm{train}} = \mathrm{diag}\{A_m^{\mathrm{T}} S_{\mathrm{w:train}} A_m\} \tag{9}$$

Since the PCA is already applied and the eigenvalues are available, within class scatter $S_{\mathrm{w:train}}$ can be easily calculated by using the well known relationship of $S_t = S_{\mathrm{b}} + S_{\mathrm{w}}$, where S_t is the total

scatter, i.e., $S_t = \frac{1}{LS}C$, C is the covariance matrix defined in Eq. (1) (Jin et al., 2001; Yang and Yang, 2003). Therefore,

$$\begin{aligned} \Lambda_{I:\text{train}} &= \text{diag}\{A_m^{\text{T}} S_{\text{w:train}} A_m\} \\ &= \Lambda - \Lambda_{E:\text{train}} \end{aligned} \tag{10}$$

where Λ is the set of eigenvalues of standard PCA corresponding to the eigenvector set A_m. With respect to the inter variation of gallery subjects, since each subject only has one face sample, the between class scatter is reduced to the total scatter, i.e.,

$$\text{Var}_{\text{inter:gallery}}(A_m) = \sum_{i=1}^{m} \Lambda_{E:\text{gallery}}(i) \tag{11}$$

$$\Lambda_{E:\text{gallery}} = \text{diag}\{A_m^{\text{T}} S_{\text{b:gallery}} A_m\} \tag{12}$$

$$S_{\text{b:gallery}} = \frac{1}{G}\sum_{i=1}^{G} (\boldsymbol{g}_i - \bar{\boldsymbol{g}})(\boldsymbol{g}_i - \bar{\boldsymbol{g}})^{\text{T}} \tag{13}$$

where $\bar{\boldsymbol{g}} = 1/G\sum_{i=1}^{G}\boldsymbol{g}_i$ is the mean of all gallery images.

4.3. Selection strategy

In order to select the optimal m combinations in A which optimize the criterion J, a proper searching strategy should be defined. In literature, many selection schemes are available with good performance for different applications, such as Sequential Forward/Backward Selection (SFS/SBS), "plus l-take away r" procedure, Sequential Floating Forward/Backward Selection (SFFS/SFBS), Branch-and-Bound, and Genetic Algorithms (GAs) (Kumar, 2000; Jain et al., 2000; Sun et al., 2004).

In this work, Sequential Forward Selection (SFS) is chosen as the selection scheme for its simplicity and good performance (Jain and Zongker, 1997). It adds features progressively. At a time, one feature is included which in combination with other previously selected features maximizes the criterion, i.e., $A(k+1) = A(k) \oplus \arg\max_{\boldsymbol{a}_i \in \mathbb{A}-A(k)} J(A(k) \oplus \boldsymbol{a}_i)$, where $A(k)$ is the selected feature set at time k, $\mathbb{A}$ is the set with all available features from which the selection is performed, a_i is the feature component and the operator $\oplus$ is used to denote the combination of two components. The procedure is repeated until the target number of features are selected.

To retain important discriminant information, the process starts with the most significant feature $\boldsymbol{a}_1$, the one corresponding to the largest eigenvalue. The whole selection procedure is summarized as follows:

(1) $A_m(1) = [\boldsymbol{a}_1]$

(2) For $k = 1$ to m

$$A_m(k+1) = A_m(k) \oplus \arg \max_{\boldsymbol{a}_i \in A - A_m(k)} J(A_m(k) \oplus \boldsymbol{a}_i)$$

With the above selection criterion, the obtained feature subspace A_m captures most of the intersubject variations while the intrasubject variations are greatly reduced, resulting in a more discriminant subspace for classification task.

Compared to the standard eigenface approach, the proposed algorithm does not increase the complexity of the FR system. The feature selection process can be applied off line during training. In the operation session, identification is performed by calculating the Euclidean distance between pairs of probe and gallery images in a feature subspace, identical to that obtained via the standard eigenface approach. The only difference is the fact that the feature subspace is spanned by the selected eigenfaces in A_m instead of the first m eigenfaces used in the traditional eigenface approach.

5. Experiments and results

5.1. Experiment setup

Realistic surveillance photo identification applications with one training sample per subject can be simulated using the FERET database. The FERET database includes 14,501 face images of 1209 subjects covering a wide range of variations in viewpoints, illuminations, facial expressions and so on. Since the focus of this letter is in recognition and not face detection, all face images are manually aligned and normalized. Thus, the coordinate information of the subjects' eyes are

considered to be available a-priori. In the current FERET database, only 3817 face images of 1200 subjects are provided with available eye coordinates information. In all experiments reported here, images are preprocessed following the FERET protocol guidelines. Namely, the following preprocessing operations are performed: (1) images are rotated and scaled so that the centers of the eyes are placed on specific pixels and the image size is normalized to 150×130; (2) a standard mask is applied to remove nonface portions; (3) histogram equalization is performed and image intensity values are normalized to zero mean and unit standard deviation; (4) each image is finally represented, after the application of mask, as a vector of dimensionality 17,154.

Among these 1200 subjects, there exist 226 subjects with 3 images per subject. These 678 images are forming the generic training set. In addition, there are 1703 images of 256 subjects with at least 4 images/subject. Of these images, 1476 are frontal images while 227 are non frontal images. These images are used to form the gallery and probe sets. We randomly select 256 frontal images (one per subject) to form the gallery set. The remaining images are considered to be the probe set. Similar experimental configuration was also suggested in (Beveridge et al., 2001). We further partition the probe set into three subsets. Set $\mathscr{P}_1$ contains 914 images of 256 subjects. The camera time difference between $\mathscr{P}_1$ probe images and their corresponding gallery matches is less than half year ($\leqslant$180 days). In $\mathscr{P}_1$, probe images are very close to their gallery matches, mostly taken in the same session. This represents the most ideal scenario, where the operating assumption is that the input images are taken in an environment similar to that of the stored templates. Set $\mathscr{P}_2$ consists of 226 images of 75 subjects. The camera time difference between the $\mathscr{P}_2$ images and their corresponding gallery matches is greater than one and half year ($\geqslant$540 days). Set $\mathscr{P}_3$ contains 227 non frontal images of 48 subjects with no particular consideration with respect to camera time. Set $\mathscr{P}_2$ includes the variations due to aging while images in $\mathscr{P}_3$ exhibit considerable pose variations, a condition often encountered when pictures are captured in an uncontrolled environment. It should be noted, that although set $\mathscr{P}_1$ represents an idealized scenario, the experiments based on sets $\mathscr{P}_2$ and $\mathscr{P}_3$ simulate the most often encountered scenarios in realistic applications such as surveillance photo identification.

Applying the PCA solution on the generic training set results in the creation of a 677-dimensional space. The cardinality M of the complete eigenface set A is determined using the 95% energy capturing rule $((\sum_{k=1}^{M}\lambda_k/\sum_{k=1}^{677}\lambda_k) > 95\%)$, resulting in a value of $M = 270$. Therefore, the complete eigenface set A consists of the first 270 eigenfaces, from which a feature subset A_m is selected according to the proposed criterion of Eq. (3) with the cardinality m up to 100.

The maximum cardinality value 100 is determined by experiment such that no performance improvement can be observed when additional features are included. Each gallery and probe image is then projected to the constructed feature space. The evaluation is performed by calculating the Euclidean distance between pairs of probe and gallery images in feature subspace spanned by the feature subset A_m:$d(j) = \|(A_m)^T(\boldsymbol{g}_i - \boldsymbol{p})\|$, where $\boldsymbol{g}_i$, $i = 1,2,\ldots,G$, is the gallery image and $\boldsymbol{p}$ is the probe. According to the FERET protocol, a probe is in the top K if the distance to its corresponding gallery match is among the K smallest distances for the gallery. Thus the recognition rate at rank K is the number of probe images in the top K divided by probe size (Phillips et al., 2000).

5.2. Results and analysis

The performance of the proposed algorithm is compared with that of the traditional eigenface scheme, Bayesian approach, LDA solution and the so-called $(PC)^2A$ method proposed in (Wu and Zhou, 2002). For the Bayesian approach, we built the intrasubject space by using generic training samples followed by a maximum likelihood matching as suggested in (Moghaddam et al., 2000). Since direct application of the LDA method on the problem under consideration, one training sample per subject, is impossible, we followed the suggestions in (Huang et al., 2003) and generated additional artificial samples by shifting the face mask by one pixel up, down, left and right. Instead of separating the face image into several local fea-

ture blocks, an approach followed (Huang et al., 2003), we treat the entire face image as a holistic pattern, greatly reducing the complexity at the expense of less than 5% reduction in the recognition rate as reported in (Huang et al., 2003).

Fig. 3 depicts the comparative evaluation of our proposed algorithm at $\eta = 0.6$ using the entire generic training set against the above mentioned approaches with different number of feature vectors, namely $N = 20$, 40, 80. Results depicted in Fig. 3 indicate that an obvious improvement can be obtained using the proposed scheme over the traditional eigenface method when probes $\mathscr{P}_2$ and $\mathscr{P}_3$ are considered. On the other end, when set $\mathscr{P}_1$ is considered, the selection procedure employed here do not result in any performance improvement. In $\mathscr{P}_1$, the time difference between probe and gallery is within half year and most of them are taken in the same session. Therefore the intra variation is considerably small compared to that in $\mathscr{P}_2$ and $\mathscr{P}_3$. In such a case, the recognition performance is mainly determined by the total intersubject variation. Although intra variation is large in the subspace, since the probe image and its gallery match are very similar, the projection of their corresponding difference $\Delta = P - G$ remains small. Therefore, in this case, intersubject variation dominates the performance. As we discussed in Section 3, the larger, in magnitude, the corresponding eigenvalue, the more inter variation is included. Thus, the traditional eigenface method performs very well. Since the proposed method retains the first eigenvector, which is believed to capture most of the intersubject variation, the performance gap between the two methods is very small.

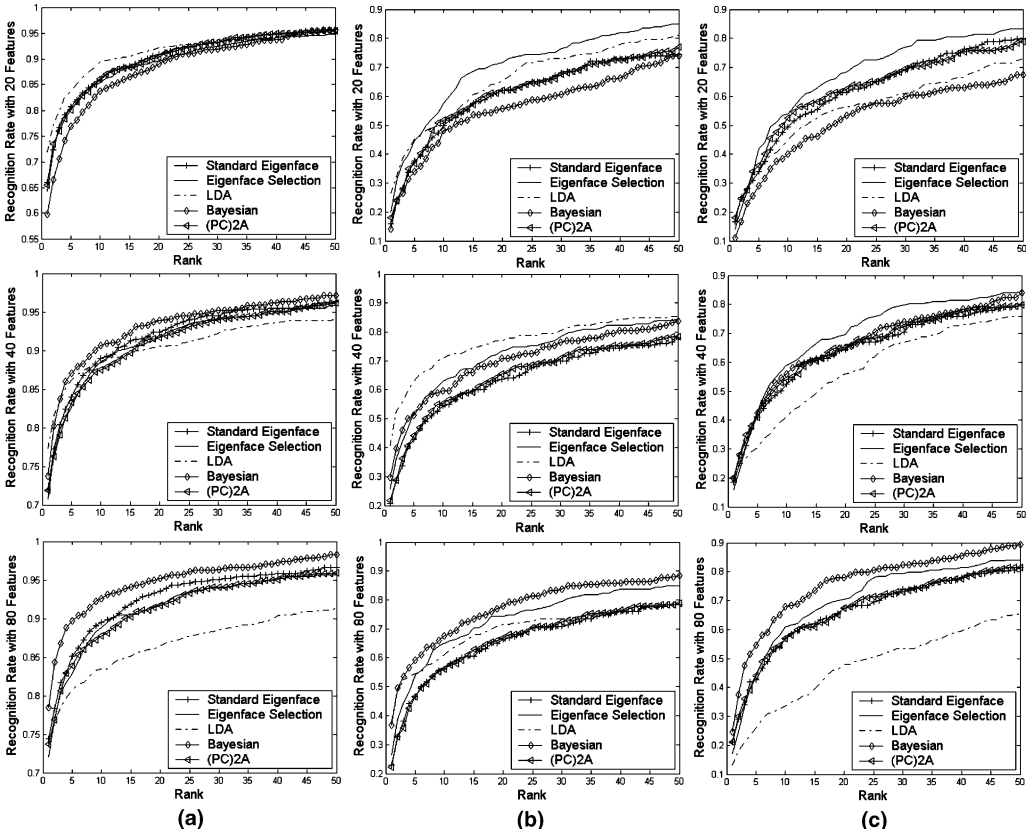

Fig. 3. Performance comparison of proposed method (eigenface selection) with $\eta = 0.6$ against standard eigenface (PCA), bayesian approach (bayesian), LDA (LDA), and $(PC)^2A$ method ((PC)2A); top: with 20 features; middle: with 40 features; bottom: with 80 features. (a) Probe 1. (b) Probe 2. (c) Probe 3.

However, in realistic applications, the probe image significantly differs from its gallery match (Δ is large), the intrasubject variation dominates the performance. In such a case, maximizing the ratio is the best way to improve recognition performance. This is the application scenario simulated using probe sets $\mathscr{P}_2$ and $\mathscr{P}_3$. In such a case, our method results in significant performance improvement compared to the one obtained using the standard PCA method.

Furthermore, the proposed method outperforms the $(PC)^2A$ approach, which reported performance similar to that of the eigenface approach. The modified LDA solution of (Huang et al., 2003) outperforms the proposed method in $\mathscr{P}_2$ when a large, in dimensionality space, representation is used (Fig. 3b, $N = 40$), however, our method returns the best results when a lower, cost effective representation is used (Fig. 3b, $N = 20$). Visual inspection of the performance curves depicted in Fig. 3 also indicates that our method outperforms the modified LDA method when $\mathscr{P}_3$ probe is considered. This is to be expected as the knowledge of pose variations was not available during the LDA-like training. As for the Bayesian scheme, the performance improves as the number of features increases. Therefore, Bayesian approach outperforms the proposed method when a large number of features are used as shown in Fig. 3, $N = 80$. However, under the condition when only small number of features are used, the most commonly encountered application where storage and processing time is of the essence, the proposed method outperforms the competition.

Fig. 4 and Tables 1 and 2 helps us to understand the effect that the regularization parameter η has on the performance of the algorithm when probe sets $\mathscr{P}_1$, $\mathscr{P}_2$ and $\mathscr{P}_3$ are considered. Fig. 4 depicts the recognition rate at rank 10 with various η, while Tables 1 and 2 list number of stored templates needed to be retrieved in order to achieve a specific recognition rate. When $\eta = 0$, intersubject variation is estimated only based on gallery images, which exactly targets at discriminating the gallery subjects. However, the performance is worse than that obtained using the generic training images ($\eta > 0$). As shown in the tables, more templates are needed to achieve a certain level of

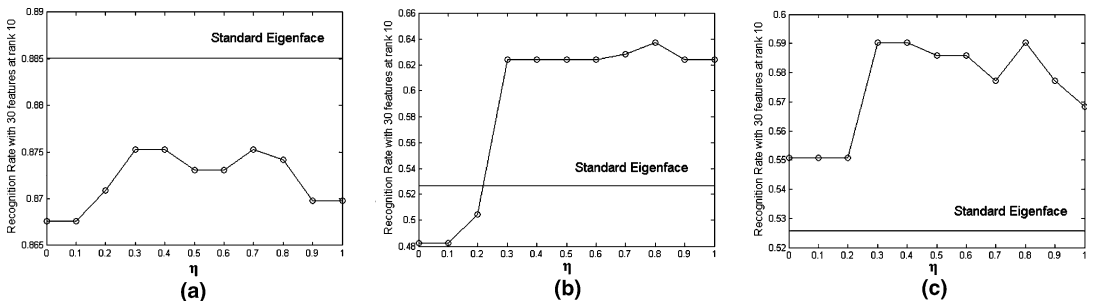

Fig. 4. Recognition rate with 30 features at rank 10 on three probe sets with varied η. (a) Probe 1. (b) Probe 2. (c) Probe 3.

Table 1
Rank of images to be extracted to achieve the following recognition rate (%) on $\mathscr{P}_1$, $\mathscr{P}_2$ and $\mathscr{P}_3$ with 20 features

η	$\mathscr{P}_1$					$\mathscr{P}_2$					$\mathscr{P}_3$				
	60	70	80	90	100	60	70	80	90	100	60	70	80	90	100
0	1	2	5	22	253	24	38	67	143	219	17	32	52	94	221
0.2	1	2	5	21	253	24	38	69	141	219	17	27	53	88	227
0.4	1	2	5	21	253	12	17	34	90	227	13	20	34	80	219
0.6	1	2	5	22	253	11	18	36	94	228	14	21	35	80	218
0.8	1	2	5	22	253	11	18	36	94	228	13	21	37	79	222
1	1	2	5	20	255	13	21	39	99	227	14	23	41	80	220

Table 2
Rank of images to be extracted to achieve the following recognition rate (%) on $\mathscr{P}_1$, $\mathscr{P}_2$ and $\mathscr{P}_3$ with 80 features

η	$\mathscr{P}_1$					$\mathscr{P}_2$					$\mathscr{P}_3$				
	60	70	80	90	100	60	70	80	90	100	60	70	80	90	100
0	1	1	4	15	254	17	31	64	135	216	14	24	46	94	220
0.2	1	1	4	15	254	17	32	64	136	216	14	24	46	95	220
0.4	1	1	3	14	254	8	17	32	86	226	10	20	30	70	215
0.6	1	1	3	12	254	8	15	31	85	224	10	19	34	72	217
0.8	1	1	3	12	254	8	15	31	83	224	10	19	33	72	216
1	1	1	4	15	254	9	16	30	90	224	12	20	32	73	219

recognition performance when $\eta = 0$ compared to that when $\eta > 0$. Fig. 4(b) suggests that the performance with $\eta = 0$ may be even worse than that of the standard eigenface approach. This confirms our claim that although reliance on the gallery images results in zero bias, it also leads to high estimation variance due to the small sample size. As η increases, with the inclusion of the generic training images, the variance of the estimation is reduced at the expense of a greater bias. The max bias value is obtained when $\eta = 1$. Therefore, as η increases, the performance firstly increases as the variance dominates the estimation error and then decreases when bias weights in. It should be reported that in the studies performed, recognition performance with $\eta = 1$, corresponding to large bias errors, is slightly worse than the best one obtained. This indicates that the instability due to small sample size is a more serious problem compared to the problems due to the bias of target subjects.

The final set of experiments deal with the quantification of the influence of the generic training set and its sample size on recognition performance. From the all generic training samples (226 subjects, 3 face images/subject), we randomly select 113 subjects with 3 face images per subject. The procedure is repeated 5 times. Therefore,

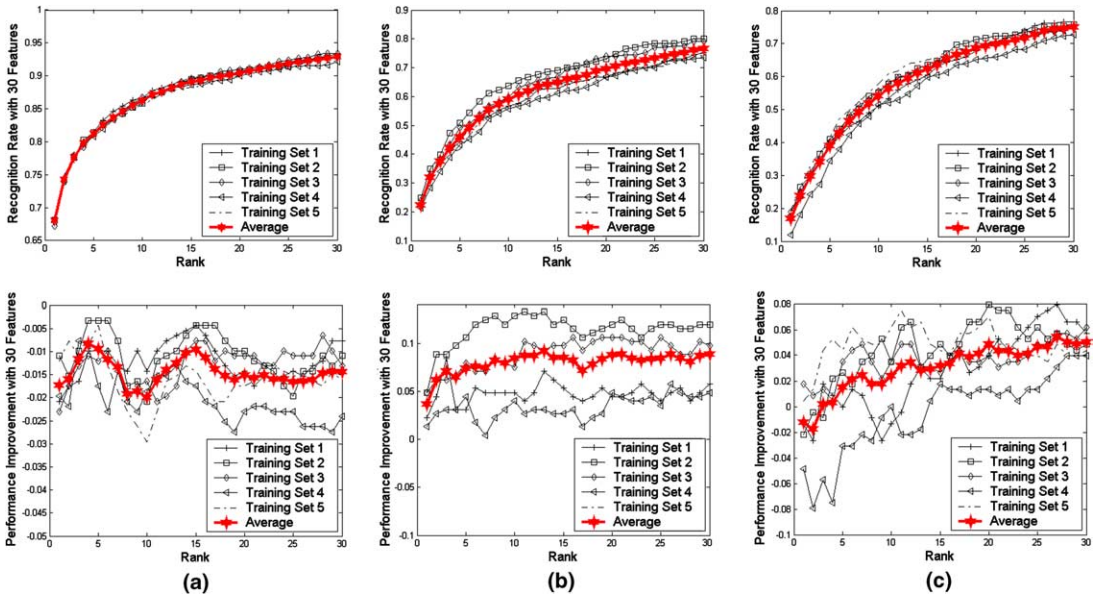

Fig. 5. Performance comparison with different generic training set on three probe sets; top: recognition rate comparison; bottom: performance improvement comparison. (a) Probe 1. (b) Probe 2. (c) Probe 3.

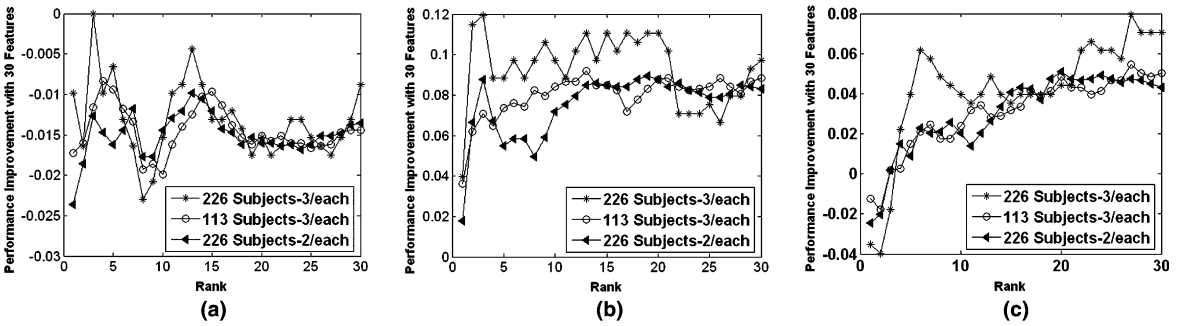

Fig. 6. Performance improve comparison with different generic training sample size on three probe sets. (a) Probe 1. (b) Probe 2. (c) Probe 3.

5 different generic training sets are generated to determine the influence of training set. Fig. 5 depicts the recognition rate at $\eta = 1$ ($R_{\text{Selection}}$) and as well as the performance improvement ($R_{\text{Selection}} - R_{\text{PCA}}$) achieved using the proposed approach over the traditional eigenface method when different training sets are used. Please note that $R_{\text{Selection}}$ and R_{PCA} refer to the recognition rate of the proposed method and eigenface approach respectively. The performance varies depending on different generic training sets used, however, the average performance for probe sets $\mathscr{P}_2$ and $\mathscr{P}_3$ is consistently better than the one reported by PCA. To determine the influence of the generic training sample size, we vary the number of the subjects as well as the number of images per subject. A comparative evaluation of performances obtained using different sample sizes (226 subjects—3 images/subject, 113 subjects—3 images/subject, 226 subjects—2 images/subject) is depicted in Fig. 6. As it can be seen, the performance improvement is proportional to the generic training sample size. As training using generic sets can be performed off line, prior to the actual applications, it is reasonable to assume that a reasonably sized generic training set is always available.

6. Conclusion

In this letter, we introduced a feature selection mechanism which can be applied to solve the problem of subject identification when only one face sample per subject is available. The proposed feature selector determines a low dimensionality feature space in which intersubject variation is maximized while intrasubject variation is minimized. Experimentation following the FERET protocol indicates that the proposed solution boosts the recognition performance, outperforming the standard eigenface approach in recognition tasks of practical importance, such as surveillance photo identification.

Acknowledgement

This work is partially supported by a grant provided by the Bell University Laboratory at the University of Toronto. The first author also acknowledges the support provided by the CITO Student Internship Program. Portions of the research in this paper use the FERET database of facial images collected under the FERET program. The authors would like to thank the FERET Technical Agent, the US National Institute of Standards and Technology (NIST) for providing the FERET database.

References

Adini, Y., Moses, Y., Ullman, S., 1997. Face recognition: The problem of compensating for changes in illumination direction. IEEE Trans. Pattern Anal Machine Intell. 19 (7), 721–732.

Bartlett, M.S., Lades, H.M., Sejnowshi, T., 2002. Face recognition by independent component analysis. IEEE Trans. Neural Networks 13 (6), 1450–1464.

Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal Machine Intell. 19 (7), 711–720.

Beveridge, J.R., She, K., Draper, B.A., Givens, G.H., 2001. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition 1, 535–542.

Brunelli, R., Poggio, T., 1993. Face recognition: Feature versus templates. IEEE Trans. Pattern Anal Machine Intell. 15 (10), 1042–1052.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. 2 (2), 121–167.

Chellappa, R., Wilson, C.L., Sirohey, S., 1995. Human and machine recognition of faces: A survey. Proc. IEEE 83, 705–740.

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification, second ed. John Wiley and Sons, New York.

Er, M.J., Wu, J., Lu, J., Toh, H.L., 2002. Face recognition with radial basis function (RBF) neural networks. IEEE Trans. Neural Networks 13 (3), 697–710.

Etemad, K., Chellappa, R., 1997. Discriminant analysis for recognition of human face images. J. Optical Soc. Amer. 14, 1724–1733.

Huang, J., Yuen, P.C., Chen, W.S., Lai, J.H., 2003. Component-based LDA method for face recognition with one training sample. IEEE Internat. Workshop Anal. Model. Faces Gestures, 120–126.

Jain, A., Zongker, D., 1997. Feature selection: Evaluation, application and small sample performance. IEEE Trans. Pattern Anal. Machine Learn. 19 (2), 153–158.

Jain, A., Robert, P.W., Mao, J., 2000. Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Machine Learn. 22 (1), 4–37.

Jin, Z., Yang, J.Y., Hu, Z.S., Lou, Z., 2001. Face recognition based on uncorrelated discriminant transformation. Pattern Recognition 34 (7), 1405–1416.

Kim, H.C., Kim, D., Bang, S.Y., 2003. Face recognition using LDA mixture model. Pattern Recognition Lett. 24 (15), 2815–2821.

Kumar, S., 2000. Modular learning through output space decomposition. Ph.D. dissertation, University of Texas, Austin.

Liu, C., Wechsler, H., 1999. Comparative assessment of independent component analysis (ICA) for face recognition. In: Proc. of the Second Internat. Conf. on Audio and Video-based Biometric Person Authentication,Washington, DC.

Lu, J., Plataniotis, K.N., Venetsanopoulos, A.N., 2003. Face recognition using LDA-based algorithms. IEEE Trans. Neural Networks 14 (1), 195–200.

Martez, A.M., Kak, A.C., 2001. PCA versus LDA. IEEE Trans. Pattern Anal. Machine Intell. 23 (2), 228–233.

Moghaddam, B., Jebara, T., Pentland, A., 2000. Bayesian face recognition. Pattern Recognition 33 (11), 1771–1782.

Perlibakas, V., 2004. Distance measures for PCA-based face recognition. Pattern Recognition Lett. 25 (6), 711–724.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P., 2000. The FERET evaluation method for face recognition algorithms. IEEE Trans. Pattern Anal. Machine Intell. 22 (10), 1090–1104.

Sun, Z., Bebis, G., Miller, R., 2004. Object detection using feature subset selection. Pattern Recognition 37 (11), 2165–2176.

Turk, M.A., Pentland, A.P., 1991. Eigenfaces for recognition. J. Cognitive Neurosci. 3 (1), 71–86.

Wang, X., Tang, X., 2003. Unified subspace analysis for face recognition. Proc. 9th IEEE Internat. Conf. on Computer Vision, 679–686.

Wu, J., Zhou, Z.H., 2002. Face recognition with one training image per person. Pattern Recognition Lett. 23 (14), 1711–1719.

Yang, J., Yang, J.Y., 2003. Why can LDA be performed in PCA transformed space. Pattern Recognition 36 (2), 563–566.

Yang, J., Zhang, D., Frangi, A.F., Yang, J.Y., 2004. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. IEEE Trans. Pattern Anal Machine Intell. 26 (1), 131–137.

Zhao, W.Y., Chellappa, R., Rosenfeld, A., Phillips, P.J., 2003. Face recognition: A literature survey. ACM Comput. Surv. 35 (4), 399–458.