Pattern Recognition Letters 33 (2012) 1157–1162

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A two-stage linear discriminant analysis for face-recognition

Alok Sharma [a,b,c,*], Kuldip K. Paliwal [a]

[a] Signal Processing Lab, Griffith University, Australia
[b] School of Engineering & Physics, University of the South Pacific, Fiji
[c] Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo, Japan

ARTICLE INFO

Article history:
Received 20 September 2011
Available online 10 February 2012
Communicated by K.A. Toh

Keywords:
Two-stage linear discriminant analysis
Small sample size problem
Classification accuracy

ABSTRACT

A two-stage linear discriminant analysis technique is proposed that utilizes both the null space and range space information of scatter matrices. The technique regularizes both the between-class scatter and within-class scatter matrices to extract the discriminant information. The regularization is conducted in parallel to give two orientation matrices. These orientation matrices are concatenated to form the final orientation matrix. The proposed technique is shown to provide better classification performance on face recognition datasets than the other techniques.

1. Introduction

Linear discriminant analysis (LDA) is a well known technique for dimensionality reduction and feature extraction (Duda et al., 2000; Sharma and Paliwal, 2006, 2008, 2010, 2012; Chen et al., 2000; Lu et al., 2003a,b, 2005; Yang et al., 2003; Yu and Yang, 2001; Swets and Weng, 1996; Belhumeur et al., 1997; Ye, 2005; Guo et al., 2007; Thomaz et al., 2005; Huang et al., 2002; Tian et al., 1986; Zhao et al., 2003; Jiang et al., 2008; Gao and Davis, 2006; Paliwal and Sharma, 2010, 2011; Mandal et al., 2010). Dimensionality reduction plays crucial role in the face recognition problem. It is generally applied for improving robustness (or generalization capability) and reducing computational complexity of the face recognition classifier. In the LDA technique, the orientation matrix $\boldsymbol{W}$ is computed from the eigenvalue decomposition (EVD) of $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$ (Duda et al., 2000), where $\boldsymbol{S}_W \in \Re^{d\times d}$ is within-class scatter matrix, $\boldsymbol{S}_B \in \Re^{d\times d}$ is between-class scatter matrix and d is the dimensionality of feature space. In the face recognition problem, the matrix $\boldsymbol{S}_W$ becomes singular and its inverse computation becomes impossible. Several techniques are reported in the literature that overcome this drawback of LDA (Chen et al., 2000; Lu et al., 2003a,b, 2005; Yang et al., 2003; Yu and Yang, 2001; Swets and Weng, 1996; Belhumeur et al., 1997; Ye, 2005; Guo et al., 2007; Thomaz et al., 2005; Sharma and Paliwal, 2010, 2012; Huang et al., 2002; Tian et al., 1986; Zhao et al., 2003; Jiang et al., 2008; Paliwal and Sharma, 2010, 2011; Mandal et al., 2010).

In LDA, there are four informative spaces namely, null space of $\boldsymbol{S}_W\left(\boldsymbol{S}_W^{null}\right)$, range space of $\boldsymbol{S}_W(\boldsymbol{S}_W^{range})$, null space of $\boldsymbol{S}_B\left(\boldsymbol{S}_B^{null}\right)$ and range space of $\boldsymbol{S}_B$ ($\boldsymbol{S}_B^{range}$). All these four individual spaces have significant discriminant information (refer Appendix I for empirical demonstration). To approximate the inverse computation of $\boldsymbol{S}_W$, different combinations of these spaces are used in the literature for finding the orientation matrix $\boldsymbol{W}$. For an instance the pseudo-inverse technique (Tian et al., 1986) uses $\boldsymbol{S}_W^{range}$ and $\boldsymbol{S}_B^{range}$ to compute the orientation matrix. The regularized LDA technique (Zhao et al., 2003) uses $\boldsymbol{S}_W^{null}$, $\boldsymbol{S}_W^{range}$ and $\boldsymbol{S}_B^{range}$. However, due to the use of small value of regularization parameter (compared to the large eigenvalues of $\boldsymbol{S}_W$), the $\boldsymbol{S}_W^{range}$ gets de-emphasize in the inverse operation of $\boldsymbol{S}_W$. Therefore, the influential spaces in the regularized LDA technique are $\boldsymbol{S}_W^{null}$ and $\boldsymbol{S}_B^{range}$. Similarly, the null LDA technique (Chen et al., 2000) uses $\boldsymbol{S}_W^{null}$ and $\boldsymbol{S}_B^{range}$. These techniques basically utilize two spaces in the orientation matrix computation and discard the other two spaces. Since the individual spaces contribute crucial discriminant information for classification, discarding some spaces would sacrifice the classification performance of the classifier. Theoretically, if all the four spaces can be inherited appropriately in the computation of orientation matrix $\boldsymbol{W}$ then the classification performance can be improved further.

In this paper, we exploit ways of utilizing all the four spaces. The inclusion of all the spaces of scatter matrices is done in two analyses. Fig. 1 illustrates the proposed strategy. The orientation matrix can be computed from the input data by carrying out two discriminant analyses in parallel. In the first analysis, the orientation matrix $\boldsymbol{W}_1$ is computed by retaining top eigenvalues and eigenvectors of $\boldsymbol{S}_W'^{-1}S_B$, where non-singular matrix $\boldsymbol{S}'$ is the approximation of singular matrix $\boldsymbol{S}$. This will retain $\boldsymbol{S}_W^{null}$ and $\boldsymbol{S}_B^{range}$. In the second analysis, the orientation matrix $\boldsymbol{W}_2$ is obtained by retaining top eigenvalues and eigenvectors of $\boldsymbol{S}_B'^{-1}\boldsymbol{S}_W$. This will retain $\boldsymbol{S}_W^{range}$ and $\boldsymbol{S}_B^{null}$. The orientation matrices obtained by these two analyses are

* Corresponding author at: Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo, Japan.

E-mail addresses: aloks@ims.u-tokyo.ac.jp, sharma_al@usp.ac.fj (A. Sharma).

0167-8655/$ - see front matter
doi:10.1016/j.patrec.2012.02.001

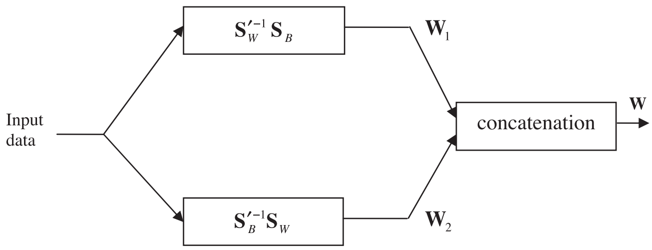

Fig. 1. The proposed strategy.

concatenated to get the final orientation matrix $\boldsymbol{W}$, i.e., $\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2]$. For brevity we call the proposed technique the two-stage LDA technique. The non-singular approximation $\boldsymbol{S}'$ of singular matrix $\boldsymbol{S}$ can be evaluated in two ways: (1) using regularized LDA technique (Zhao et al., 2003) where $\boldsymbol{S}' = \boldsymbol{S} + \alpha \boldsymbol{I}$ (α is the regularization parameter); and, (2) using extrapolation technique (Jiang et al., 2008; Sharma and Paliwal, 2010) where eigenvalues of $\boldsymbol{S}$ are extrapolated by applying curve fitting or some criterion function. In this paper we show that the resulting orientation matrix $\boldsymbol{W}$ provides better classification results than other existing techniques.

2. Notations and descriptions

Let us denote the n linearly independent training samples (or feature vectors) in d-dimensional space by $\mathcal{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, having class labels $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, where $\omega_i \in \{1, 2, \ldots, c\}$ and c is the number of classes. The set $\mathcal{X}$ can be subdivided into c subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_c$ where each subset $\mathcal{X}_j$ belongs to a particular class label and consists of n_j number of samples such that:

$$n = \sum_{j=1}^{c} n_j$$

and $\mathcal{X}_j \subset \mathcal{X}$ and $\mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_c = \mathcal{X}$.

Let $\boldsymbol{\mu}_j$ be the centroid of $\mathcal{X}_j$ and $\boldsymbol{\mu}$ be the centroid of $\mathcal{X}$, then the between class scatter matrix $\boldsymbol{S}_B$, within-class scatter matrix $\boldsymbol{S}_W$ and total-scatter matrix $\boldsymbol{S}_T$ are defined as (Duda et al., 2000)

$$\boldsymbol{S}_B = \sum_{j=1}^{c} \boldsymbol{n}_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{S}_W = \sum_{j=1}^{c} \boldsymbol{S}_j \tag{2}$$

where

$$\boldsymbol{S}_j = \sum_{\boldsymbol{x} \in \mathcal{X}_j} (\boldsymbol{x} - \boldsymbol{\mu}_j)(\boldsymbol{x} - \boldsymbol{\mu}_j)^{\mathrm{T}}$$

and

$$\boldsymbol{S}_T = \sum_{\boldsymbol{x} \in \mathcal{X}} (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \tag{3}$$

Since in the face recognition task $d > n$, the scatter matrices $\boldsymbol{S}_B$, $\boldsymbol{S}_W$ and $\boldsymbol{S}_T$ will be singular with ranks $r_b = c - 1$, $r_w = n - c$ and $r_t = n - 1$, respectively. The null space of $\boldsymbol{S}_T$ carries no discriminative information (Huang et al., 2002), therefore, the dimensionality can be reduced from d-dimensional space to $r_t = n - 1$ dimensional space by applying principal component analysis (PCA) as a pre-processing step to remove the null space of $\boldsymbol{S}_T$. This would make the technique computationally faster. The range space of total scatter matrix $\boldsymbol{U}_{TR} \in \Re^{d \times r_t}$ will be used as a transformation. This will give us transformed within-class scatter matrix $\widehat{\boldsymbol{S}}_W \in \Re^{r_t \times r_t}$ and transformed between-class scatter matrix $\widehat{\boldsymbol{S}}_B \in \Re^{r_t \times r_t}$. These matrices can be decomposed as

$$\widehat{\boldsymbol{S}}_W = \boldsymbol{U}_W \boldsymbol{D}_W^2 \boldsymbol{U}_W^{\mathrm{T}} \tag{4}$$

and

$$\widehat{\boldsymbol{S}}_B = \boldsymbol{U}_B \boldsymbol{D}_B^2 \boldsymbol{U}_B^{\mathrm{T}} \tag{5}$$

where $\boldsymbol{D}_W \in \Re^{r_t \times r_t}$ and $\boldsymbol{D}_B \in \Re^{r_t \times r_t}$ are diagonal matrices whose elements (arranged in descending order) are the square-root of the eigenvalues of $\widehat{\boldsymbol{S}}_W$ and $\widehat{\boldsymbol{S}}_B$, respectively; and $\boldsymbol{U}_W \in \Re^{r_t \times r_t}$ and $\boldsymbol{U}_B \in \Re^{r_t \times r_t}$ are orthogonal matrices consisting of the corresponding eigenvectors as columns. Since the rank of $\widehat{\boldsymbol{S}}_W$ is r_w, the matrix $\boldsymbol{U}_W$ can be formed as $\boldsymbol{U}_W = [\boldsymbol{U}_{WR}, \boldsymbol{U}_{WN}]$ where $\boldsymbol{U}_{WR} \in \Re^{r_t \times r_w}$ corresponds to the range space of $\widehat{\boldsymbol{S}}_W$ and $\boldsymbol{U}_{WN} \in \Re^{r_t \times (r_t - r_w)}$ corresponds to the null space of $\widehat{\boldsymbol{S}}_W$. In a similar way, we can write $\boldsymbol{U}_B = [\boldsymbol{U}_{BR}, \boldsymbol{U}_{BN}]$ where $\boldsymbol{U}_{BR} \in \Re^{r_t \times r_b}$ corresponds to the range space of $\widehat{\boldsymbol{S}}_B$ and $\boldsymbol{U}_{BN} \in \Re^{r_t \times (r_t - r_b)}$ corresponds to the null space of $\widehat{\boldsymbol{S}}_B$.

3. Two-stage LDA technique

It is well known in the literature that the null space of $\widehat{S}_W$ contains crucial information for classification (Chen et al., 2000; Ye, 2005). The null space based LDA techniques retain the null space information of $\widehat{S}_W$, however, they discard the range space information of $\widehat{\boldsymbol{S}}_W$. It has been seen that the range space information of $\widehat{\boldsymbol{S}}_W$ is also important for classification (Swets and Weng, 1996; Belhumeur et al., 1997) and by discarding it could penalize classification performance. Some techniques (e.g. Guo et al., 2007; Zhao et al., 2003; Jiang et al., 2008; Sharma and Paliwal, 2010) estimates non-singular within-class scatter matrix $\widehat{\boldsymbol{S}}'_W$ by adding a small positive constant (known as regularization parameter) to the eigenvalues of $\widehat{\boldsymbol{S}}_W$ (Guo et al., 2007; Zhao et al., 2003) or by extrapolating the eigenvalues of $\widehat{\boldsymbol{S}}_W$ in its null space (Jiang et al., 2008; Sharma and Paliwal, 2010). Thereafter, obtaining the eigenvectors corresponding to the top eigenvalues of $\widehat{\boldsymbol{S}}'^{-1}_W \widehat{\boldsymbol{S}}_B$. In these techniques the null space information of $\widehat{\boldsymbol{S}}_W$ and the range space information of $\widehat{\boldsymbol{S}}_B$ are effectively retained. Although, the range space information of $\widehat{\boldsymbol{S}}_W$ is utilized in these techniques, it has very less influence as it is de-emphasized in the inverse operation of $\widehat{\boldsymbol{S}}_W$ (see Fig. 2). Nonetheless, theoretically the latter implementation would contain more information than the former techniques. To see the qualitative contribution of $\widehat{\boldsymbol{S}}'^{-1}_W \widehat{\boldsymbol{S}}_B$ in obtaining the orientation matrix, we decompose $\widehat{\boldsymbol{S}}'_W$ into its eigenvalues and eigenvectors as

$$\widehat{\boldsymbol{S}}'_W = \boldsymbol{U}_W \widehat{\boldsymbol{D}}_W^2 \boldsymbol{U}_W^{\mathrm{T}} \tag{6}$$

where diagonal matrix $\widehat{\boldsymbol{D}}_W = \begin{bmatrix} \Sigma_W & 0 \\ 0 & \widehat{\Sigma}_W \end{bmatrix}$, $\Sigma_W \in \Re^{r_w \times r_w}$ and $\widehat{\Sigma}_W \in \Re^{(r_t - r_w) \times (r_t - r_w)}$ is the estimation or regularization of eigenvalues Σ_W.

From Eq. (5), $\widehat{\boldsymbol{S}}_B$ can be formed as

$$\widehat{\boldsymbol{S}}_B = [\boldsymbol{U}_{BR}, \boldsymbol{U}_{BN}] \begin{bmatrix} \Sigma_B^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{BR}^{\mathrm{T}} \\ \boldsymbol{U}_{BN}^{\mathrm{T}} \end{bmatrix} = \boldsymbol{U}_{BR} \Sigma_B^2 \boldsymbol{U}_{BR}^{\mathrm{T}} \tag{7}$$

where $\Sigma_B \in \Re^{r_b \times r_b}$.

From Eqs. (6) and (7), we can write

$$\widehat{\boldsymbol{S}}'^{-1}_W \widehat{\boldsymbol{S}}_B = \boldsymbol{U}_W \widehat{\boldsymbol{D}}_W^{-2} \boldsymbol{U}_W^{\mathrm{T}} \boldsymbol{U}_{BR} \Sigma_B^2 \boldsymbol{U}_{BR}^{\mathrm{T}} \tag{8}$$

The EVD of Eq. (8) can be computed and the range space information of $\widehat{\boldsymbol{S}}'^{-1}_W \widehat{\boldsymbol{S}}_B$ can be used in the formation of orientation matrix. Three things can be observed here:

(1) The null space of $\widehat{\boldsymbol{S}}_B$ is discarded.
(2) The range space information of within-class scatter matrix in the inverse operation is de-emphasized.
(3) The null space of the product $\widehat{\boldsymbol{S}}'^{-1}_W \widehat{\boldsymbol{S}}_B$ is discarded.

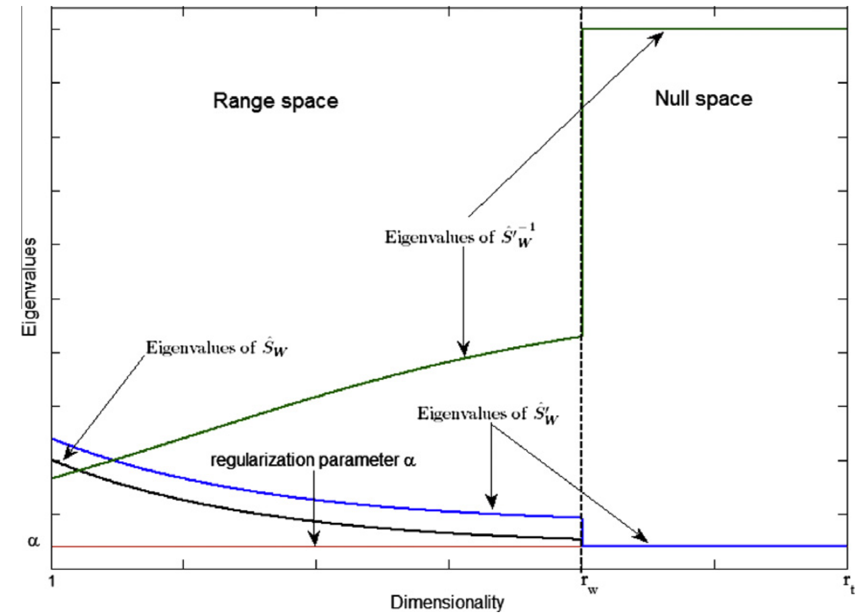

Fig. 2. This figure uses regularization method to get non-singular estimate $\widehat{\boldsymbol{S}}'_W$ from the singular matrix $\widehat{\boldsymbol{S}}_W$ and illustrates the de-emphasis of the range space information of $\widehat{\boldsymbol{S}}_W$ in its inverse operation. The terms r_w and r_t are the ranks of $\widehat{\boldsymbol{S}}_W$ and $\widehat{\boldsymbol{S}}_T$, respectively. The region between 1 and r_w is the range space of $\widehat{\boldsymbol{S}}_W$ and the region between r_w and r_t is the null space of $\widehat{\boldsymbol{S}}_W$. The eigenvalues of $\widehat{\boldsymbol{S}}_W$ are added by the regularization parameter α which gives the eigenvalues of $\widehat{\boldsymbol{S}}'_W$ (i.e., $\widehat{\boldsymbol{S}}'_W = \widehat{\boldsymbol{S}}_W + \alpha I$). The regularization parameter is usually a small quantity obtained by performing cross-validation procedure on the training feature vectors. This parameter addition helps in defining the eigenvalues of $\widehat{\boldsymbol{S}}'_W$ in the null space region. Thereby enabling the inverse operation of $\widehat{\boldsymbol{S}}'_W$. The small eigenvalues of $\widehat{\boldsymbol{S}}'_W$ (in the null space) get emphasized in the inverse operation. These eigenvalues are used as weighting coefficients for their corresponding eigenvectors and therefore the eigenvectors of $\widehat{\boldsymbol{S}}'_W$ in the range space are de-emphasized.

It is known that though the null space of $\widehat{\boldsymbol{S}}_B$ is less effective, it contains some useful information for classification (Gao and Davis, 2006; Paliwal and Sharma, 2010, Appendix I). Therefore, theoretically if the null space of $\widehat{\boldsymbol{S}}_B$ is included in computing the orientation matrix then the classification performance can be improved. Furthermore, if the range space of $\widehat{\boldsymbol{S}}_W$ can be utilized effectively then it can help in retaining more information. Next, if the eigenvectors of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$ are represented by $\boldsymbol{E} = [\boldsymbol{E}_R, \boldsymbol{E}_I]$ (where $\boldsymbol{E}_R \in \Re^{r_t \times r_b}$ and $\boldsymbol{E}_I \in \Re^{r_t \times (r_t - r_b)}$) then it is possible that some eigenvalues (which are not in the range space of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$) are complex valued which would give complex eigenvectors as columns of $\boldsymbol{E}_I \in \Re^{r_t \times (r_t - r_b)}$ and cannot be included in the formation of orientation matrix. Some of the eigenvalues of a singular matrix can become complex due to limited size of the hardware (Golub and Loan, 1996). Since the matrix $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$ is positive semi-definite and singular (with rank r_b), then theoretically it should produce r_b positive eigenvalues and the remaining eigenvalues should be zero. However, due to the hardware limitations, it may produce some very small non-zero eigenvalues (positive or negative). The small negative eigenvalues will lead to complex eigenvectors. For example, if the size of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$ is 10×10 and its rank is 3 then it will give 3 eigenvectors corresponding to the positive eigenvalues which are defined as its range space $\boldsymbol{E}_R$. The remaining 7 eigenvectors define the null space $\boldsymbol{E}_I$, some of its eigenvectors corresponding to very small negative eigenvalues will be complex valued. In our implementation we use only the $\boldsymbol{E}_R$ and discard $\boldsymbol{E}_I$.

In order to retain more information for the purpose of improving the classification performance further, we investigate a strategy to: (1) include the null space of $\widehat{\boldsymbol{S}}_B$, (2) include the range space of $\widehat{\boldsymbol{S}}_W$, and (3) extract the null space information of $\widehat{\boldsymbol{S}}'^{-1}_W\boldsymbol{S}_B$.

One strategy would be to estimate eigenvalues for the null space of $\widehat{\boldsymbol{S}}_B$ (as done e.g. for $\widehat{\boldsymbol{S}}_W$ in regularized LDA technique) and perform eigenvalue decomposition of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}'_B$. The term $\widehat{\boldsymbol{S}}'_B$ is the regularized or estimated matrix of $\widehat{\boldsymbol{S}}_B$ and can be defined as

$$\widehat{\boldsymbol{S}}'_B = \boldsymbol{U}_B \widehat{\boldsymbol{D}}^2_B \boldsymbol{U}^{\mathrm{T}}_B \tag{9}$$

where $\widehat{\boldsymbol{D}}_B = \begin{bmatrix} \Sigma_B & 0 \\ 0 & \widehat{\Sigma}_B \end{bmatrix}$ and $\widehat{\Sigma}^{(r_t - r_b)\times(r_t - r_b)}_B$ is the estimation of eigenvalues in the null space of $\widehat{\boldsymbol{S}}_B$. This strategy may satisfy above points 1 and 3. However, it could have either no effect on classification performance or can deteriorate the classification performance. See Appendix II for details.

In order to satisfy the above three points we can do as follows. Consider a matrix $\boldsymbol{C}$ which has leading and lagging eigenvectors represented by $\boldsymbol{L}$ and $\boldsymbol{G}$, respectively. Then the leading and lagging eigenvectors of $\boldsymbol{C}^{-1}$ can be given by $\boldsymbol{G}$ and $\boldsymbol{L}$, respectively. Therefore, $\widehat{\boldsymbol{S}}_B$ can be estimated to be non-singular matrix $\widehat{\boldsymbol{S}}'_B$ to retain its null space[1] which can be used to approximate the null space of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$ by obtaining the range space of $\widehat{\boldsymbol{S}}'^{-1}_B\widehat{\boldsymbol{S}}_W$. The eigenvectors of $\widehat{\boldsymbol{S}}'^{-1}_B\widehat{\boldsymbol{S}}_W$ can be denoted by $\widehat{\boldsymbol{E}} = [\widehat{\boldsymbol{E}}_R, \widehat{\boldsymbol{E}}_I]$ (where $\widehat{\boldsymbol{E}}_R \in \Re^{r_t \times r_w}$ and $\widehat{\boldsymbol{E}}_I \in \Re^{r_t \times (r_t - r_w)}$). Since the rank of $\widehat{\boldsymbol{S}}_B$ is $r_b < r_w$, only leading r_b eigenvectors (i.e., eigenvectors corresponding to largest eigenvalues) of $\widehat{\boldsymbol{E}}_R$ can be considered to form an orientation matrix. The remaining $r_w - r_b$ eigenvalues could be noisy which would give erroneous corresponding weighted eigenvectors. If the leading eigenvectors of $\widehat{\boldsymbol{E}}_R$ is denoted by $\widehat{\boldsymbol{E}}_{RL} \in \Re^{r_t \times r_b}$ then it can be considered as approximated null space of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$. Since the range space of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$ is $\boldsymbol{E}_R$ and its approximated null space is $\widehat{\boldsymbol{E}}_{RL}$, the orientation matrix in r_t-dimensional space would be $\widehat{\boldsymbol{W}} = [\boldsymbol{E}_R, \widehat{\boldsymbol{E}}_{RL}] \in \Re^{r_t \times 2r_b}$ or in d-dimensional space would be $\boldsymbol{W} = \boldsymbol{U}_{TR}\widehat{\boldsymbol{W}} \in \Re^{d \times 2r_b}$. Theoretically, this strategy would include all the four spaces and retrieve the null space information of $\widehat{\boldsymbol{S}}'^{-1}_W\widehat{\boldsymbol{S}}_B$. The summary of the algorithm is depicted in Table 1.

4. Computational considerations

The computational complexity of the two-stage LDA technique is higher than other techniques like null space based technique

[1] Regularization of $\widehat{\boldsymbol{S}}_B$ can be done in a similar manner as we have done regularization of $\widehat{\boldsymbol{S}}_W$ matrix. An example of this can be viewed from Fig. 2 by replacing the matrix $\widehat{\boldsymbol{S}}_W$ by matrix $\widehat{\boldsymbol{S}}_B$ and by replacing the rank r_w by rank r_b.

Table 1
The algorithm.

Step 1.	Pre-processing stage: apply PCA to find range space $\boldsymbol{U}_{TR} \in \Re^{d \times r_t}$ of total scatter matrix $\boldsymbol{S}_T$ and apply it to find transformed within-class scatter matrix $\widehat{\boldsymbol{S}}_W \in \Re^{r_t \times r_t}$ and between class scatter matrix $\widehat{\boldsymbol{S}}_B \in \Re^{r_t \times r_t}$ (where r_t is the rank of $\boldsymbol{S}_T$)
Step 2.	Estimate non-singular matrices $\widehat{\boldsymbol{S}}'_W$ and $\widehat{\boldsymbol{S}}'_B$ from singular matrices $\widehat{\boldsymbol{S}}_W$ and $\widehat{\boldsymbol{S}}_B$, respectively, by using either regularization technique or extrapolation technique
Step 3.	Decompose $\widehat{\boldsymbol{S}}'^{-1}_W \widehat{\boldsymbol{S}}_B$ into its eigenvalues and eigenvectors, and find the leading r_b number of eigenvectors $\widehat{\boldsymbol{W}}_1 \in \Re^{r_t \times r_b}$ (i.e., eigenvectors corresponding to largest eigenvalues), where r_b is the rank of between-class scatter matrix
Step 4.	Similarly (as Step 3) find leading r_b number of eigenvectors $\widehat{\boldsymbol{W}}_2 \in \Re^{r_t \times r_b}$ from the eigenvalue decompose of $\widehat{\boldsymbol{S}}'^{-1}_B \widehat{\boldsymbol{S}}_W$
Step 5.	Form $\widehat{\boldsymbol{W}} = [\widehat{\boldsymbol{W}}_1, \widehat{\boldsymbol{W}}_2]$ and compute orientation matrix $\boldsymbol{W} = \boldsymbol{U}_{TR}\widehat{\boldsymbol{W}} \in \Re^{d \times 2r_b}$

(OLDA) (Ye, 2005), PCA plus LDA (Swets and Weng, 1996; Belhumeur et al., 1997) and DLDA (Yu and Yang, 2001) as eigenvector computation is required both in the null space as well as in the range space of scatter matrices. However, by applying PCA as a pre-processing step the computational complexity would be reduced by removing the null space of total scatter matrix and then transforming feature vectors on the r_t-dimensional space. The computational complexity of the pre-processing step (Step 1) would be $O(dn^2)$, where d is the dimensionality of feature vectors and n is the number of training feature vectors. The computational complexity of eigenvalue decomposition of the scatter matrices in Step 2 would be around $O(n^3)$. Some additional computational complexity will be required to estimate eigenvalues in the null space of scatter matrices depending upon the technique used. The computational complexity of Steps 3 and 4 would be $O(n^3)$ and of Step 5 would be $O(dn^2)$.

5. Experimental setup and results

Four commonly known datasets namely ORL database (Samaria and Harter, 1994), AR database (Martinez, 2002), Yale (Belhumeur) and FERET (Phillips et al., 2000) are utilized for the experimentation. The ORL database contains 400 images of 40 people having 10 images per subject. The dimensionality of the original feature space is 10304. The AR database contains 100 classes. We use a subset of AR database with 14 face images per subject. The dimensionality is 4980. The Yale database contains 165 images of 15 subjects. There are 11 images per subject, one for each of the following facial expressions or configurations: center-light, with glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised and wink (Belhumeur). All the images are first cropped to 65×51, therefore, the dimensionality is 3315. For FERET database, we used 6 images per subject. In total, 85 subjects are utilized for the experiment. The image is first cropped to 84×64; i.e., the dimensionality is 5796. The proposed technique is compared with the following techniques: DLDA (Yu and Yang, 2001), PCA plus LDA technique (Swets and Weng, 1996; Belhumeur et al., 1997), null space based technique (OLDA) (Ye, 2005), regularized LDA technique (Zhao et al., 2003) (the regularization parameter was estimated by using leave-one out cross-validation procedure on training set.), regularized LDA based on DLDA framework (the η parameter was estimated by using leave-one out cross validation procedure on training set) (Lu et al., 2005), maximum uncertainty LDA (MLDA) technique (Thomaz et al., 2005) and eigenfeature regularization technique (Jiang et al., 2008). Table 2 shows the average recognition accuracy on four datasets using all the techniques. For the two-stage LDA technique, we use here the extrapolation technique (Sharma and Paliwal, 2010; Jiang et al., 2008) for computing non-singular estimates of scatter matrices. The results for regularization technique (Zhao et al., 2003) are given later in this section. The nearest neighbor classifier using Euclidean distance measure is used for classifying a test vector. Multiple runs (5 runs) of N-fold cross-validation are applied to find the average recognition accuracy, where $N = 2$. We can see from Table 2 that the two-stage LDA technique outperforms the other techniques.

Table 2
Performance of the techniques in terms of average recognition accuracy over multiple runs of N-fold cross-validation on ORL, AR, Yale and FERET databases.

Techniques	ORL (%)	AR (%)	Yale (%)	FERET (%)
DLDA (Yu and Yang, 2001)	89.5	80.8	93.5	92.9
Null space based technique (OLDA) (Ye, 2005)	91.5	80.8	97.3	97.1
PCA plus LDA (Swets and Weng, 1996; Belhumeur et al., 1997)	86.0	83.4	98.0	95.7
Regularized LDA (Zhao et al., 2003)	91.5	75.4	97.9	97.3
Regularized LDA based on DLDA framework (Lu et al., 2005)	89.8	81.6	94.7	94.5
MLDA (Thomaz et al., 2005)	92.0	76.2	97.9	97.8
Eigenfeature regularization technique (Jiang et al., 2008)	92.3	81.8	98.6	97.7
Two-stage LDA technique	92.6	87.8	98.7	98.0

We have already shown in Appendix I that null space of $\boldsymbol{S}_B$ contains useful information for classification. In order to show whether this space provides additional and complementary information over the three other spaces (range space of $\boldsymbol{S}_W$, null space of $\boldsymbol{S}_W$ and range space of $\boldsymbol{S}_B$), we report here the results for the two-stage LDA technique with and without the null space of $\boldsymbol{S}_B$. For the two-stage LDA technique with the null space of $\boldsymbol{S}_B$, the procedure is same as described in Table 1. For using the two-stage LDA technique without the null space of $\boldsymbol{S}_B$, we modified the procedure given in Table 1 as follows. Instead of using the range space of $\widehat{\boldsymbol{S}}'^{-1}_B \widehat{\boldsymbol{S}}_W$, we use the range space of $\widehat{\boldsymbol{S}}_W$ in Step 4. Multiple runs of N-fold cross-validation are carried out on all the four datasets and the resulting average recognition accuracies with the null space of $\boldsymbol{S}_B$ and without the null space of $\boldsymbol{S}_B$ are depicted in Table 3. It can be observed from this table that the null space of $\boldsymbol{S}_B$ does provide complementary information over the other three spaces and plays a useful role in the proposed two-stage LDA technique.

So far we have provided results where the two-stage LDA technique is used to reduce the dimensionality to $2r_b$. Now we show its performance as a function of dimensionality. To demonstrate this, we varied the dimensions from 5 to $2r_b$ (where, $2r_b = 2(c-1)$ and c is the number of classes) and computed the average recognition accuracy by doing multiple runs of N-fold cross-validation technique (as above). The results are demonstrated in Table 4. It can be seen from this table that the recognition performance improves by increasing the dimensionality.

In the experiments described above, we have used the extrapolation procedure for obtaining the non-singular estimates of scatter matrices in the two-stage LDA technique. Now we use the regularization method to obtain the non-singular estimate of these matrices. In order to do this, we vary the regularization parameter α in the following manner. For estimating within-class scatter matrix in full space we define regularization parameter $\alpha = \delta * \lambda_W$, where λ_W is the maximum eigenvalue of within-class scatter matrix and δ is a small positive number. Similarly, for estimating between-class scatter matrix in full space we define $\alpha = \delta * \lambda_B$, where λ_B is the maximum eigenvalue of between-class scatter matrix. The average recognition accuracy is then obtained by conducting multiple runs of N-fold cross-validation on the four face recognition datasets. The results are shown in Table 5. We can see from this table that the recognition performance can be improved by choosing the regularization parameter appropriately. However, it must be

Table 3
Average classification accuracy with and without the null space of $\boldsymbol{S}_B$.

Datasets	With null space of $\boldsymbol{S}_B$	Without null space of $\boldsymbol{S}_B$
ORL (%)	92.6	90.6
AR (%)	87.8	71.0
Yale (%)	98.7	91.5
FERET (%)	98.0	93.6

Table 4
Recognition performance as a function of number of features (dimensions).

Dataset	Dimensions/number of features							
	5	10	20	25	50	100	r_b	$2r_b$
ORL (%)	81.0	86.7	89.2	89.6	92.3	–	91.9 (r_b = 39)	92.6 ($2r_b$ = 78)
AR (%)	26.6	47.3	63.2	67.0	77.8	86.9	87.0 (r_b = 99)	87.8 ($2r_b$ = 198)
Yale (%)	84.1	95.1	98.8	98.8	–	–	98.8 (r_b = 14)	98.7 ($2r_b$ = 28)
FERET (%)	46.7	67.3	84.3	87.5	94.2	97.3	97.1 (r_b = 84)	98.0 ($2r_b$ = 168)

Table 5
Recognition performance by varying the value of regulation parameter.

Datasets	δ = 0.001	δ = 0.1	δ = 0.3	δ = 0.5	δ = 0.7	δ = 0.9	δ = 1.0
ORL (%)	86.5	87.5	90.3	90.8	90.8	91.0	90.8
AR (%)	62.9	74.4	72.7	71.7	71.0	70.4	70.3
Yale (%)	94.3	93.9	93.3	91.9	90.8	89.7	89.7
FERET (%)	91.6	95.5	95.7	94.3	93.9	93.7	93.7

Table A1
Classification accuracy using $\boldsymbol{S}_W^{null}$, $\boldsymbol{S}_W^{range}$, $\boldsymbol{S}_B^{range}$ and $\boldsymbol{S}_B^{null}$.

Dataset	$\boldsymbol{S}_W^{null}$(%)	$\boldsymbol{S}_W^{range}$(%)	$\boldsymbol{S}_B^{range}$(%)	$\boldsymbol{S}_B^{null}$(%)
ORL (Samaria and Harter, 1994)	91.0	87.0	88.5	43.0
AR (Martinez, 2002)	81.6	70.1	70.4	41.0
Yale (Belhumeur)	98.9	84.4	92.2	63.3
FERET (Phillips et al., 2000)	96.9	90.6	94.1	62.4

noticed (by comparing Tables 4 and 5) that the performance of the two-stage LDA technique with extrapolation procedure is in general better than that with the regularization procedure.

6. Conclusion

We have proposed a two-stage LDA technique that includes both the null space and range space information of between-class scatter and within-class scatter matrices. The regularization is done in parallel to give two orientation matrices. These orientation matrices are concatenated to form the final orientation matrix. The proposed technique is shown to provide better classification performance on several face recognition datasets than the other techniques.

Appendix I

In this appendix we describe (pragmatically) that all the four spaces namely, null space of $\boldsymbol{S}_W\left(\boldsymbol{S}_W^{null}\right)$, range space of $\boldsymbol{S}_W\,(\boldsymbol{S}_W^{range})$, null space of $\boldsymbol{S}_B\left(\boldsymbol{S}_B^{null}\right)$ and range space of $\boldsymbol{S}_B\,(S_B^{range})$ contain information for discriminant analysis. In order to demonstrate this, first we project the original feature vectors onto the range space of total scatter matrix as a pre-processing step. Then all the spaces are utilized individually to do dimensionality reduction and to classify a test feature vector, the nearest neighbor classifier is used. For this experiment the datasets have been approximately equally divided into training samples and test sample. Table A1 depicts the classification accuracy. It can be observed from the table that individual spaces ($\boldsymbol{S}_W^{null}, \boldsymbol{S}_W^{range}$ and $\boldsymbol{S}_B^{range}$) contain significant discriminant information. Though the $\boldsymbol{S}_B^{null}$ is less effective, it still contains some information.

Appendix II

In this appendix we will show that by doing eigenvalue decomposition of $\widehat{\boldsymbol{S}}_W'^{-1}\widehat{\boldsymbol{S}}_B'$ could result in noisy eigenvectors (where the full rank scatter matrices $\widehat{\boldsymbol{S}}_W'$ and $\widehat{\boldsymbol{S}}_B'$ are the estimates of singular scatter matrices $\widehat{\boldsymbol{S}}_W$ and $\widehat{\boldsymbol{S}}_B$, respectively). From Eqs. (6) and (9) of the text, $\boldsymbol{Q} = \widehat{\boldsymbol{S}}_W'^{-1}\widehat{\boldsymbol{S}}_B'$ can be expressed as

$$\begin{aligned}\boldsymbol{Q} &= [\boldsymbol{U}_{WR}, \boldsymbol{U}_{WN}]\begin{bmatrix}\Sigma_W^{-2} & 0\\ 0 & \widehat{\Sigma}_W^{-2}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_{WR}^{\mathrm{T}}\\ \boldsymbol{U}_{WN}^{\mathrm{T}}\end{bmatrix}[\boldsymbol{U}_{BR}, \boldsymbol{U}_{BN}]\begin{bmatrix}\Sigma_B^{2} & 0\\ 0 & \widehat{\Sigma}_B^{2}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_{BR}^{\mathrm{T}}\\ \boldsymbol{U}_{BN}^{\mathrm{T}}\end{bmatrix}\\ &= \boldsymbol{P} + \left(\boldsymbol{U}_{WN}\widehat{\Sigma}_W^{-2}\boldsymbol{U}_{WN}^{\mathrm{T}}\right)\left(\boldsymbol{U}_{BN}\widehat{\Sigma}_B^{2}\boldsymbol{U}_{BN}^{\mathrm{T}}\right)\end{aligned}$$

where $\boldsymbol{P}$ is the remaining sum of products. If the diagonal entries of $\widehat{\Sigma}_W$ is $\hat{\lambda}_j > 0$ (for $j = 1,\ldots,(r_t - r_w)$) and $\widehat{\Sigma}_B$ is $\hat{\gamma}_k > 0$ (for $j = 1,\ldots,(r_t - r_b)$), and the corresponding column vectors of $\boldsymbol{U}_{WN}$ is $\boldsymbol{u}_j$ and $\boldsymbol{U}_{BN}$ is $\boldsymbol{v}_k$ then

$$= \boldsymbol{P} + \left(\sum_{j=1}^{r_t-r_w}\boldsymbol{u}_j\boldsymbol{u}_j^{\mathrm{T}}/\hat{\lambda}_j^2\right)\left(\sum_{k=1}^{r_t-r_b}(\boldsymbol{v}_k\boldsymbol{v}_k^{\mathrm{T}})\hat{\gamma}_k^2\right) \tag{A1}$$

Since $\hat{\lambda}_j^2$ and $\hat{\gamma}_k^2$ are lagging eigenvalues of $\widehat{\boldsymbol{S}}_W'$ and $\widehat{\boldsymbol{S}}_B'$, respectively, the eigenvalues will be small and noisy. It is reasonable to assume that the values of $\hat{\lambda}_j$ (for all j) are closely equal and similarly the values of $\hat{\gamma}_k$ (for all k) are closely equal; i.e., $\hat{\lambda}_1 \approx \hat{\lambda}_2 \cdots \approx \hat{\lambda}_{r_t-r_w}$ and $\hat{\gamma}_1 \approx \hat{\gamma}_2 \cdots \approx \hat{\gamma}_{r_t-r_b}$. Therefore, Eq. (A1) can be written as

$$\approx \boldsymbol{P} + \frac{\hat{\gamma}^2}{\hat{\lambda}^2}\left(\sum_{j=1}^{r_t-r_w}\boldsymbol{u}_j\boldsymbol{u}_j^{\mathrm{T}}\right)\left(\sum_{k=1}^{r_t-r_b}\boldsymbol{v}_k\boldsymbol{v}_k^{\mathrm{T}}\right) \quad \text{(the subscript of eigenvalues are removed)}$$

Let the eigenvalue $\hat{\gamma}$ consists of true eigenvalue $\hat{\gamma}_0$ and additive noise σ_b, where $|\sigma_b| \leqslant b$ and b is a positive constant. Similarly, let $\hat{\lambda} = \lambda_0 + \sigma_w$, where $|\sigma_w| \leqslant w$ and w is a positive constant. Let ε denotes the ratio $\hat{\gamma}^2/\hat{\lambda}^2$. The ratio ε could be in the range $0 < \varepsilon < \infty$ and if noise σ_b and σ_w are dominant factors then this will lead to serious erroneous value of $\boldsymbol{Q}$ and the orientation matrix.

References

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.

Belhumeur, P.N. <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>.

Chen, L.-F., Liao, H.-Y.M., Ko, M.-T., Lin, J.-C., Yu, G.-J., 2000. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition 33, 1713–1726.

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification. Wiley, New York.

Gao, H., Davis, J.W., 2006. Why direct LDA is not equivalent to LDA. Pattern Recognition 39, 1002–1006.

Golub, G.H., Loan, C.F.V., 1996. Matrix Computations. The John Hopkins University Press.

Guo, Y., Hastie, T., Tinshirani, R., 2007. Regularized discriminant analysis and its application in microarrays. Biostatistics 8 (1), 86–100.

Huang, R., Liu, Q., Lu, H., Ma, S. 2002. Solving the small sample size problem of LDA. In: Proceedings of ICPR, vol. 3, 2002, pp. 29–32.

Jiang, X., Mandal, B., Kot, A., 2008. Eigenfeature regularization and extraction in face recognition. IEEE Trans. Pattern Anal. Machine Intell. 30 (3), 383–394.

Lu, J., Plataniotis, K.N., Venetsanopoulos, A.N., 2003a. Regularized discriminant analysis for the small sample problem in face recognition. Pattern Recognition Lett. 24, 3079–3087.

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

Lu, J., Plataniotis, K.N., Venetsanopoulos, A.N., 2005. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. Pattern Recosnition Lett. 26, 181–191.

Mandal, B., Jiang, X., Eng, H.-L., Kot, A., 2010. Prediction of eigenvalues and regularization of eigenfeatures for human face verification. Pattern Recognition Lett. 31, 717–724.

Martinez, A.M., 2002. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Anal. Machine Intell. 24 (6), 748–763.

Paliwal, K.K., Sharma, A., 2010. Improved direct LDA and its application to DNA microarray gene expression data. Pattern Recognition Lett. 31 (16), 2489–2492.

Paliwal, K.K., Sharma, A., 2011. Approximate LDA technique for dimensionality reduction in small sample size case. J. Pattern Recognition Res. 6 (2), 298–306.

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

Samaria, F., Harter, A., 1994. Parameterization of a stochastic model for human face identification. In: Proceedings of the Second IEEE Workshop Applications of Computer Vision, pp. 138–142.

Sharma, A., Paliwal, K.K., 2006. Class-dependent PCA, LDA and MDC: a combined classifier for pattern classification. Pattern Recognition 39 (7), 1215–1229.

Sharma, A., Paliwal, K.K., 2008. Rotational linear discriminant analysis for dimensionality reduction. IEEE Trans. Knowl. Data Eng. 20 (10), 1336–1347.

Sharma, A., Paliwal, K.K., 2010. Regularisation of eigenfeatures by extrapolation of scatter-matrix in face-recognition problem. Electron. Lett. 46 (10), 682–683.

Sharma, A., Paliwal, K.K., 2012. A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices. Pattern Recognition 45 (6), 2205–2213.

Swets, D.L., Weng, J., 1996. Using discriminative eigenfeatures for image retrieval. IEEE Trans. Pattern Anal. Machine Intell. 18 (8), 831–836.

Thomaz, C.E., Kitani, E.C., Gillies, D.F., 2005. A maximum uncertainty LDA-based approach for limited sample size problems – with application to face recognition. In: Proceedings of the 18th Brazilian Symposium on Computer Graphics and Image Processing. IEEE CS Press, pp. 89–96.

Tian, Q., Barbero, M., Gu, Z.H., Lee, S.H., 1986. Image classification by the Foley–Sammon transform. Opt. Eng. 25 (7), 834–840.

Yang, J., Zhang, D., Yang, J.-Y., 2003. A generalised K–L expansion method which can deal with small samples size and high-dimensional problems. Pattern Anal. Appl. 6, 47–54.

Ye, J., 2005. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. J. Machine Learn. Res. 6, 483–502.

Yu, H., Yang, J., 2001. A direct LDA algorithm for high-dimensional data-with application to face recognition. Pattern Recognition 34, 2067–2070.

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