SPECTRAL ANALYSIS FOR FACE RECOGNITION

Xiaofei He†, Shuicheng Yan‡, Yuxiao Hu, Haifeng Li* and Hong-Jiang Zhang

Microsoft Research Asia, Beijing 100080, China
†Department of Computer Science, University of Chicago
‡School of Mathematical Sciences, Peking University, China
*Department of Computer Science, University of Toronto, Canada

ABSTRACT

Different eigenspace-based approaches have been proposed for the recognition of faces, i.e. eigenface, fisherface and Laplacianface. For fisherfaces, the original image space is reduced to an \( n-c \) dimensional subspace in which the standard LDA is carried out, where \( n \) is the number of training samples and \( c \) is the number of classes. In this paper, we present a spectral analysis of fisherface which shows that the initial PCA dimensionality reduction might be insufficient. The noise might not be completely eliminated. This is due to the fact that fisherface only takes into account the discriminating structure while ignores geometrical structure. Based on the theoretical analysis, we propose a new method, called enhanced fisherface, which takes into account the discriminating structure as well as the intrinsic geometrical structure. Experimental results show that the proposed approach is effective in improving the performance of face recognition.

1. INTRODUCTION

In the last decade, eigenspace-based approaches have attracted much attention in face recognition [1][2][6][9][14][16]. When using eigenspace-based methods, a face image is usually represented as a point in the image space (given by the number of pixels in the image). However, the dimensionality of the image space is typically too large to allow robust and fast face recognition. Learnability thus necessitates dimensionality reduction. Among various dimensionality reduction techniques, the linear techniques are especially attractive due to their simplicity and efficiency. In particular, Principal Component Analysis (PCA) [16], Linear Discriminant Analysis (LDA) [1], and Locality Preserving Projections (LPP) [2][4] have been applied to face recognition with impressive results.

PCA and LPP aim to discover the geometrical structure of the data points. PCA projects the data along the directions of maximal variance. What it sees is the global Euclidean structure. LPP projects the data along the directions which preserve the neighborhood structure. LPP is obtained by finding the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold; hence what it sees is the local manifold structure. Different from PCA and LPP which encode geometrical information, LDA encodes discriminating information. LDA searches for the projection axes on which the data points of different classes are far from each other and at the same time where the data points of a same class are close to each other. While PCA is an unsupervised learning algorithm and LDA is a supervised learning algorithm, it is important to note that LPP can be performed either in supervised manner or unsupervised manner.

When applying LDA and LPP to face recognition, one needs to solve a generalized eigenvector problem. Specifically, it has the following form:

\[
Ax = \lambda Bx
\]

Unfortunately, in face recognition problems, \( B \) is always singular [1][2]. For LDA, this stems from the fact that the rank of \( B \) is at most \( n-c \), where \( n \) is the number of training samples and \( c \) is the number of classes. To deal with this difficulty, Belhumeur et. al.[1] proposed the fisherfaces approach that first projects the face images into a \( n-c \) dimensional subspace obtained by PCA and LDA is carried out in such a low dimensional subspace.

While fisherface takes into account the discriminating structure, it ignores the geometrical structure of the data points. Let us consider an extreme case as follows. The training set has only two classes. For each class, the difference between any two face image vectors is a constant vector. Thus, the data points have only two degrees of freedom, and hence the rank of the matrix \( B \) in fisherfaces is only 4. This tells us that when the data points are highly correlated, the intrinsic dimensionality of the face subspace might be much lower than \( n-c \). Due to this consideration, we proposed a new approach, called enhanced...
fisherface, which takes into account the geometrical structure of the data points.

The reminder of this paper is organized into following sections. In section 2, we give a spectral analysis of fisherface method. In section 3, we propose a new method called enhanced fisherface to face recognition based on LDA. Some computational issues are discussed in section 4. The experimental results on face recognition are presented in section 5. Finally, we give concluding remarks and future work in section 6.

2. SPECTRAL ANALYSIS OF FISHERFACE

In recent years, spectral techniques have been widely used in various areas, such as clustering [11][13][15], dimensionality reduction [10], image segmentation [12], image database analysis [17], etc. In this section, we give a theoretical analysis for fisherface method using spectral techniques.

Fisherface method makes use of Linear Discriminant Analysis (LDA) to discover the discriminating structure of the data set. LDA seeks directions that are efficient for discrimination. Suppose we have a set of n d-dimensional samples \( x_1, x_2, \ldots, x_n \), belonging to l classes of faces. The objective function is as follows,

\[
\max_{w} \frac{w^{T}S_{B}w}{w^{T}S_{W}w}
\]

where \( S_{B} \) is the between-class scatter matrix and \( S_{W} \) is the within-class scatter matrix.

By simple algebra formulation, we can rewrite the matrix \( S_{W} \) as follows:

\[
S_{W} = \sum_{i=1}^{l} n_i E \left[ (x^{(i)} - m^{(i)})(x^{(i)} - m^{(i)})^{T} \right]
\]

where \( m \) is the total sample mean vector, \( n_i \) is the number of samples in the \( i \)-th class, \( m^{(i)} \) are the average vectors of the \( i \)-th class, and \( x^{(i)} \) are the sample vectors associated to the \( i \)-th class. We call \( S_{B} \) the within-class scatter matrix and \( S_{W} \) the between-class scatter matrix.

Thus, we get:

\[
S_{W} = nXWX^{T}
\]

It is interesting to note that we could regard the matrix \( W \) as the weight matrix of a graph with data points as its nodes. Specifically, \( W_{ij} \) is the weight of the edge \( (x_i, x_j) \). The matrix \( L \) is thus called graph Laplacian [3] (or, Laplace Matrix), which plays key role in LPP [4].

Similarly, we can compute the matrix \( S_{B} \) as follows:

\[
S_{B} = \sum_{i=1}^{l} n_i \left[ (m^{(i)} - \bar{m})(m^{(i)} - \bar{m})^{T} \right]
\]

where \( X = [x_1, x_2, \ldots, x_n] \) is the data covariance matrix of the \( i \)-th class. \( X = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}] \) is a \( d \times n_i \) matrix whose column vectors belong to the \( i \)-th class, and \( L = I - 1/n \cdot \mathbf{e} \mathbf{e}^{T} \) is an \( n \times n \) matrix where \( I \) is the identity matrix and \( \mathbf{e} = (1,1,\ldots,1)^{T} \) is a \( n \)-dimensional vector. To further simplify the above equation, we define:

\[
W_{ij} = \begin{cases} 1/n_i & \text{if } x_i \text{ and } x_j \text{ both belong to the } k^{th} \text{ class} \\ 0 & \text{otherwise} \end{cases}
\]

Thus, we get:

\[
S_{W} = XLX^{T}
\]

Theorem: The rank of \( L \) is \( n - c \).

Prove: Without loss of generality, we assume that the data points are ordered according to which class they are in, so that \( \{x_1, \ldots, x_n\} \) are in the first class, \( \{x_{n_1+1}, \ldots, x_{n_1+n_2}\} \)
are in the second class, etc. Thus, we can write $L$ as follows:

$$
L = \begin{bmatrix}
L_1 & L_2 & \cdots & L_c
\end{bmatrix}
$$

where $L_i$ is a square matrix,

$$
L_i = \begin{bmatrix}
1 - \frac{1}{n_i} & -\frac{1}{n_i} & \cdots & -\frac{1}{n_i} \\
-\frac{1}{n_i} & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & -\frac{1}{n_i} \\
-\frac{1}{n_i} & \cdots & -\frac{1}{n_i} & 1 - \frac{1}{n_i}
\end{bmatrix}
$$

By adding all but the first column vectors to the first column vector and then subtracting the first row vector from any other row vectors, we get the following matrix:

$$
\begin{bmatrix}
0 & 0 & \cdots & 0 \\
0 & 1 & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & 1
\end{bmatrix}
$$

Therefore, the rank of $L_i$ is $n_i - 1$ and hence the rank of $L$ is $n - c$.

Theorem 1 tells us that the rank of $XLX^T$ is at most $n - c$. However, in many cases in appearance-based face recognition, the number of pixels in an image (or, the dimensionality of the image space) is larger than $n - c$, i.e., $d > n - c$. Thus, $XLX^T$ is singular. In order to overcome the complication of a singular $XLX^T$, Belhumeur et al. [1] proposed the fisherface approach that the face images are projected from the original image space to a subspace with dimensionality $n - c$.

One problem of the fisherface approach is that, the data points might be highly correlated as we described in the introduction section and hence the rank of $X$ might be very low, even lower than $n - c$. In this case, even though we project the data points into an $n - c$ dimensional PCA subspace, the matrix $XLX^T$ might be still singular and the noise might not be completely eliminated.

In order to overcome the above problem of the conventional fisherface approach, two kinds of information need to be properly processed,

1. Discriminating information which is contained in the matrix $L$.
2. Geometrical information which is contained in the data matrix $X$.

In the next section, we will describe a new method to face recognition which takes into account the geometrical structure of the face space and is less sensitive to noise.

### 3. OUR METHOD

Based on the theoretical analysis presented in the previous section, we propose a new method to face recognition which applies PCA and LDA.

The algorithmic procedure can be formally stated below:

1. **PCA projection:** We compute the eigenvectors and eigenvalues of the data covariance matrix. Let $\lambda_{PCA}^1 \geq \lambda_{PCA}^2 \geq \cdots \geq \lambda_{PCA}^d$ denote the $d$ eigenvalues. Suppose $\lambda_1, \lambda_2, \ldots, \lambda_k$ contains 98% information in the sense of reconstruction such that,

$$
\sum_{i=1}^k \lambda_i \approx 98\%
$$

We project the image set $\{x_i\}$ into the PCA subspace with dimensionality $\min \{k, n - c\}$. We denote the transformation matrix of PCA by $W_{PCA}$.

2. **LDA projection:** In the PCA subspace, we apply the standard LDA to reduce the dimension to $c - 1$.

More formally, we solve the following eigenvector problem:

$$
Sw = \lambda S_w w
$$

or,

$$
Cw = (1 + \lambda)XLX^T w
$$

Let $w_1, w_2, \ldots, w_{c-1}$ be the solutions of the above equation, ordered according to their eigenvalues, $\lambda_{LDA}^1 \geq \lambda_{LDA}^2 \geq \cdots \geq \lambda_{LDA}^{c-1}$. Thus, the optimal projection is given by:

$$
W = W_{LDA}W_{PCA}
$$

$$
W_{LDA} = [w_1, w_2, \ldots, w_{c-1}]
$$

3. **Recognition:** The training samples are projected into a $c - 1$ dimensional subspace by $W$. When a new image comes, it is also projected into such a subspace by $W$. Finally, the new image is identified by a nearest-neighbor classifier.

It is important to note that in the first step the data points are projected into a PCA subspace such that,

1. The dimensionality of the subspace is no larger than the rank of $L$. The rank of the matrix $X$ equals to the dimensionality of the subspace. Hence the matrix $S_w$ (or, $XLX^T$) has full rank.
2. The data points in the PCA subspace are uncorrelated * . The noise is eliminated. Here, the noise corresponds to the smallest eigenvalues of PCA. Note that, in fisherface, the noise might not be eliminated due to the fact that fisherface ignores the geometrical structure of the face space.

4. COMPUTATIONAL ISSUES

In PCA subspace, we need to compute equation (1). Suppose the dimensionality of the PCA subspace is m. Because \( S_W \) has full rank in the subspace, equation (1) can be simply converted to ordinary eigenvector problem as follow:

\[
S_W^{-1} S_B w = \lambda w
\]  

(2)

However, in most cases \( S_W^{-1} S_B \) is not symmetric. Hence the computation of equation (2) is unstable. In our experiments, we apply Singular Value Decomposition (SVD) to achieve a stable computation. It is easy to check that \( S_W \) is symmetric and positive definite. We first decompose it as follows:

\[
S_W = U S U^T
\]

where \( S = diag\{s_1, s_2, ..., s_m\} \) is an diagonal matrix whose entries are the eigenvalues of \( S_W \) and \( U \) is an orthonormal matrix whose column vectors are the eigenvectors of \( S_W \) and \( U U^T = U^T U = I \). Because \( S_W \) is positive definite, \( s_i > 0, i = 1, ..., m \). Thus, we can define

\[
S^{1/2} = diag\{s_1^{1/2}, s_2^{1/2}, ..., s_m^{1/2}\}
\]

\[
S^{-1/2} = diag\{s_1^{-1/2}, s_2^{-1/2}, ..., s_m^{-1/2}\}
\]

By simple algebra formulation, we have:

\[
S_W^{1/2} S_B^{1/2} = \lambda S_W^{1/2} U S_U^{1/2} U^T w
\]

Define,

\[
V = S_W^{1/2} U S_U^{1/2} U^T w
\]

Finally, the equation (1) is reduced to,

\[
V b = \lambda b
\]

(3)

The solutions to equation (1) are given by:

\[
w = U S_U^{1/2} b
\]

It is easy to check that \( V \) is symmetric and positive semi-definite, and hence the computation of equation (3) is more stable than that of equation (1).

5. EXPERIMENTAL RESULTS

In this section, several experiments are performed to show the effectiveness of our proposed method for face recognition.

5.1 Data Preparation

In this study, two face databases were tested. The first one is the Yale database [18], and the second one is the MSRA database collected at Microsoft Research Asia. In all the experiments, preprocessing to locate the faces was applied. Original images were normalized (in scale and orientation) such that the two eyes were aligned at the same position. Then, the facial areas were cropped into the final image for matching. The size of each cropped image in all the experiments is 32×32 pixels, with 256 gray levels per pixel. Thus, each image can be represented by a 1024-dimensional vector in image space.

* In PCA subspace, each dimension corresponds to a random variable and the data correlation corresponds to the correlation of the random variables. The data points are uncorrelated means that the random variables are pairwisely uncorrelated. In other words, the data correlation matrix in the PCA subspace is a diagonal matrix.
further preprocessing is done. Figure 5 shows an example of the original face image and the cropped image.

For each database, the face images are randomly divided into two sets. One is for training, and the other is for testing. Different pattern classifiers have been applied for face recognition, including nearest-neighbor [16], Bayesian [5][7], and support vector machine [8], etc. In this paper, we apply nearest-neighbor classifier for its simplicity.

5.2 Face Recognition on YALE Database

The Yale face database [18] is constructed at the Yale Center for Computational Vision and Control. It contains 165 grayscale images of 15 individuals. The images demonstrate variations in lighting condition (left-light, center-light, right-light), facial expression (normal, happy, sad, sleepy, surprised, and wink), and with/without glasses.

The face subspace is constructed by our method to best preserve the discriminating structure and the geometrical structure. For each image, it can be projected into the face subspace by the transformation matrix $W$. Since there are 15 individuals, the face subspace has 14 (=15-1) dimensions.

Several tests were performed. In each test, the database was randomly split into two subsets, training set and testing set. The training sets contain various numbers of images, from four to eight. We also applied “Leave One Out” (LOO) strategy to test the face recognition performance. The recognition was carried out by using the conventional fisherface approach and our proposed enhanced fisherface approach. Table 1 shows the recognition results. It is found that our new method outperforms the conventional fisherface method.

Figure 3 demonstrates the energy contained in the top $n$ ($n = 1, \ldots, 90$) eigenvalues obtained in the third test (90 for training and 75 for testing). As can be seen, 98% energy is contained in the top 48 eigenvalues. In conventional fisherface, the original image space (with 256 dimensions) is only reduced to an $n-c$ (= 90-15 = 75) dimensional face subspace. Therefore, there might be still some noise in such a face subspace.

5.3 Face Recognition on MSRA Database

The MSRA database was collected at Microsoft Research Asia. It contains 12 individuals, captured in two different sessions with different backgrounds and illuminations. 64 to 80 face images are collected for each individual in each session. All the faces are frontal. Figure 6 shows the sample cropped face images from this database.

For the sake of simplicity, let $MSRA-S1$ denote the image set taken in the first session, and $MSRA-S2$ denote the image set taken in the second session. $MSRA-S2$ was exclusively used for testing. 10%, 40%, 70%, and 100% face images were randomly selected from $MSRA-S1$ for training. For each of the former three cases, we repeated 20 times and computed the error rate in average. Since there are 12 individuals, the face subspace has 11 dimensions. The experimental results were shown in Table 2.

As can be seen, our method takes advantage of more training samples. The performance of our method improves significantly as the number of training sample increases. The recognition rate increases from 70.83% with 10% training samples to 86.32 with 100% training samples. There is no convincing evidence for fisherface approach that it can take advantage of more training samples. It does not perform better as the number of training samples increases. On the contrary, the performance of fisherface with only 10% training samples (73.04% recognition rate) is even better than that with 70% training samples (70.57% recognition rate). This shows that our new method is more robust and efficient for face recognition.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a new method to face recognition which takes into account not only the discriminating structure but also the geometrical structure. Theoretical analysis shows that the within-class scatter matrix $S_w$ can be represented as a product of three matrices, i.e. $S_w = X L X^T$. The matrix $L$ contains discriminating information while $X$ contains geometrical information. Based on the theoretical analysis, we propose to project the data points into a PCA subspace in which $S_w$ is guaranteed to be of full rank and the data points are uncorrelated. Moreover, the noise corresponding to the smallest eigenvalues is eliminated. Experiment results show that the new method is superior to the conventional fisherface method.

Theoretical analysis shows that Linear Discriminant Analysis has strong connections to Principal Component Analysis and Locality Preserving Projections [4] via the data covariance matrix and the Laplace matrix ($L$). What LDA sees is the discriminating structure, while PCA sees global geometrical structure and LPP sees local geometrical structure. It seems promising to combine these three kinds of information. But it remains unclear how to combine them in a principled manner.
Table 1. Performance comparison on the YALE database

<table>
<thead>
<tr>
<th></th>
<th>training set</th>
<th>testing set</th>
<th>Fisherface (error rate)</th>
<th>Our method (error rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>7</td>
<td>80.00</td>
<td>84.76</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>82.22</td>
<td>81.11</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>80</td>
<td>82.66</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>86.67</td>
<td>85.00</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>84.44</td>
<td>86.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leave one out</td>
<td>93.33</td>
<td>93.94</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Performance comparison on the MSRA database

<table>
<thead>
<tr>
<th>Percentage of MSRA-A used for training</th>
<th>Fisherface (error rate)</th>
<th>Our method (error rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>73.04</td>
<td>70.83</td>
</tr>
<tr>
<td>40%</td>
<td>79.94</td>
<td>85.15</td>
</tr>
<tr>
<td>70%</td>
<td>70.57</td>
<td>84.23</td>
</tr>
<tr>
<td>100%</td>
<td>73.57</td>
<td>86.32</td>
</tr>
</tbody>
</table>

Figure 3. Percentage of energy vs. number of accumulated eigenvalues. As can be seen, 98% energy is contained in the top 48 eigenvalues, while in fisherface the original image space is reduced to a $n-c (=90-15=75)$ dimensional subspace which still has noise.

7. REFERENCES