Fast communication

Gabor feature-based face recognition using supervised locality preserving projection

Zhonglong Zheng\textsuperscript{a,*}, Fan Yang\textsuperscript{a}, Wenan Tan\textsuperscript{a}, Jiong Jia\textsuperscript{a}, Jie Yang\textsuperscript{b}

\textsuperscript{a}Institute of Information Science and Engineering, 143# mailbox, Zhejiang Normal University, Zhejiang 321004, China
\textsuperscript{b}Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China

Received 11 June 2006; received in revised form 21 February 2007; accepted 11 March 2007
Available online 30 March 2007

Abstract

This paper introduces a novel Gabor-based supervised locality preserving projection (GSLPP) method for face recognition. Locality preserving projection (LPP) is a recently proposed method for unsupervised linear dimensionality reduction. LPP seeks to preserve the local structure which is usually more significant than the global structure preserved by principal component analysis (PCA) and linear discriminant analysis (LDA). In this paper, we investigate its extension, called supervised locality preserving projection (SLPP), using class labels of data points to enhance its discriminant power in their mapping into a low-dimensional space. The GSLPP method, which is robust to variations of illumination and facial expression, applies the SLPP to an augmented Gabor feature vector derived from the Gabor wavelet representation of face images. We performed comparative experiments of various face recognition schemes, including the proposed GSLPP method, PCA method, LDA method, LPP method, the combination of Gabor and PCA method (GPCA) and the combination of Gabor and LDA method (GLDA). Experimental results on AR database and CMU PIE database show superior of the novel GSLPP method.

\textcopyright 2007 Elsevier B.V. All rights reserved.

Keywords: Supervised learning; Face recognition; Manifold learning; Gabor wavelets; Locality preserving projection

1. Introduction

Face recognition may be applied to a wide range of fields from security systems to virtual reality systems. Therefore, it has attracted much attention of worldwide researchers. As face images are distributed in a very high-dimensional space, it is impossible to carry out research work in the face image space directly. Thus, it is essential to reduce the higher-dimensional space to a lower-dimensional space which is usually called feature space.

In the past decades, there have been many methods proposed for dimensionality reduction [1–8]. Two canonical forms of them are principal component analysis (PCA) and multidimensional scaling (MDS). Both of them are eigenvector methods aimed at modeling linear variability in the multidimensional space. PCA computes the linear projections of the greatest variance from the top eigenvectors of the data covariance matrix. MDS, however, computes the low-dimensional embedding that best preserves pairwise distances...
between data points. The results of MDS will be equivalent to PCA if the similarity is Euclidean distance. Both methods are simple to implement and not prone to local minima. However, for the data on a nonlinear sub-manifold embedded in the feature space, the results given by PCA preserve only the global structure. In many cases, local structure is emphasized especially when using nearest neighbor classifier.

Locally linear embedding and Laplacian Eigenmap are nonlinear local approaches proposed recently to discover the nonlinear structure of the manifold [3,6,7]. The essence of the two methods is seeking to map nearby points on a manifold to nearby points in a low-dimensional space. Isomap is a nonlinear global approach based on MDS and seeks to preserve the intrinsic geometry of the data [5]. These nonlinear methods have achieved impressive results both on some benchmark artificial and some real applications [9–14]. Nevertheless, the nonlinearity makes them computationally expensive. In addition, the mappings derived from them are defined on the training set and how to evaluate a novel test data remains unclear.

Recently, an unsupervised linear dimensionality reduction method, locality preserving projection (LPP), was proposed and applied to real data sets [10–13]. LPP aims to preserve the local structure of the multidimensional structure instead of global structure preserved by PCA. In addition, LPP shares some similar properties compared with LLE such as a locality preserving character. However, their objective functions are totally different. LPP is the optimal linear approximation to the eigenfunctions of the Laplace Beltrami operator on the manifold [15]. LPP is linear and can deal with new data easily. In contrast, LLE is nonlinear and unclear on how to evaluate test points.

In this paper, we describe a supervised variant of LPP, called the supervised locality preserving projection (SLPP) algorithm. Unlike LPP, SLPP projects high-dimensional data to the embedded low space taking class membership relations into account. This allows obtaining well-separated clusters in the embedded space. It is worthwhile to highlight the discriminant power of SLPP by using class information besides inheriting the properties of LPP. Therefore, SLPP demonstrates powerful recognition performance when applied to some pattern recognition tasks. The Gabor-based supervised locality preserving projection (GSLPP) method for face recognition, which is robust to variations of illumination and facial expression, applies the SLPP to an augmented Gabor feature vector derived from the Gabor wavelet representation of face images. We performed comparative experiments of various face recognition schemes, including the proposed GSLPP method, PCA method, linear discriminant analysis (LDA) method, LPP method, the combination of Gabor and PCA method (GPCA) and the combination of Gabor and LDA method (GLDA).

The rest of the paper is organized as follows: Section 2 describes LPP vs. PCA and LDA. The proposed SLPP is described in Section 3. A brief introduction to Gabor wavelets is given in Section 4. In Section 5, we apply the proposed GSLPP to some real face data sets to test its performance compared with PCA, LDA, LPP, GPCA and GLDA. Finally, we provide some concluding remarks and suggestions for future work in Section 6.

2. LPP vs. PCA and LDA

More formally, let us consider a set of $M$ sample images taking values in an $n$-dimensional image space $X = \{x_1, \ldots, x_M\}$, and assume that each image belongs to one of $c$ classes $X_1, \ldots, X_c$. PCA seeks a linear transformation mapping the original-dimensional image space into an $r$-dimensional feature space, where $r < n$. Then the transformed new feature vectors $y_k \in \mathbb{R}^r$ are defined as follows:

$$y_k = W^T x_k. \quad (1)$$

The total scatter matrix of original sample images $S_T$ is defined as

$$S_T = \sum_{k=1}^{M} (x_k - \mu)(x_k - \mu)^T, \quad (2)$$

where $M$ is the number of sample images and $\mu$ is the mean image of all samples. Then the scatter of the transformed feature vectors $Y = \{y_1, \ldots, y_M\}$ is $W^T S_T W$. The optimal projection $W_{\text{PCA}}$ of PCA is selected to maximize the determinant of the total scatter matrix of the transformed samples [16], i.e.:

$$W_{\text{PCA}} = \arg \max_w \{W^T S_T W\}. \quad (3)$$

While PCA searches for efficient directions of representation, LDA seeks efficient direction of discrimination. The between-class scatter matrix and the within-class scatter matrix are defined as
follows, respectively:

\[
S_B = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T,
\]

\[
S_W = \sum_{i=1}^{c} \sum_{j \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T,
\]

where \(N_i\) is the number of samples in class \(X_i\), and \(\mu_i\) is the mean image of class \(X_i\). The optimal projection \(W_{\text{LDA}}\) of LDA is chosen as the matrix with orthonormal columns which maximizes the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples, i.e.

\[
W_{\text{LDA}} = \arg \max_w \frac{|W^T S_B W|}{|W^T S_W W|}.
\]

We can conclude that both PCA and LDA aim to preserve the global structure. In fact, the local structure is more important in many real cases. LPP is a linear approximation algorithm of the nonlinear Laplacian eigenmap for learning a locality preserving subspace [15]. LPP aims to preserve the intrinsic geometry of the data and local structure. The objective function of LPP is defined as

\[
\min \sum_y \|y_i - y_j\|^2 S_{ij},
\]

where \(S\) is a symmetry similarity measure matrix. A possible way of defining such \(S\) is

\[
S_{ij} = \begin{cases} 
\exp(-\|x_i - x_j\|^2/t), & \|x_i - x_j\|^2 < \varepsilon, \\
0 & \text{otherwise},
\end{cases}
\]

where \(\varepsilon > 0\) defines the radius of the local neighborhood, \(t \in \mathbb{R}\). Here the selection of \(\varepsilon\) is somewhat like that of in local linear embedding (LLE) algorithm [6]. The imposed constraint is \(y^T D y = 1\) [10]. Finally, the minimization problem reduces to the following form:

\[
\arg \min W^T X L X^T W
\]

with

\[
W^T X D X^T W = 1,
\]

where \(D\) is a diagonal matrix, \(D_{ii} = \sum_j S_{ij}\). And \(L = D - S\) is the Laplacian matrix [3]. The transformation vector \(W_{\text{LPP}}\) is determined by the minimum eigenvalue solution to the generalize eigenvalue problem:

\[
X L X^T W = \lambda X D X^T W.
\]

For more detailed information about LPP, please refer to [10–12,15].

### 3. Supervised locality preserving projection

From the above analysis, both PCA and LPP are unsupervised learning methods. They do not take the class membership relation into account. Improbably speaking, one of the differences between them lies in the global or local preserving property. The locality preserving property leads to the fact that LPP outperforms PCA [10,13,15]. While PCA and LDA are all global methods, LDA utilizes the class information to enhance its discriminant ability. That is the reason why LDA outperforms PCA [16–20]. To some extent, we can say that both locality preserving property and discriminant ability are significant in learning a new feature subspace.

It is our motivation to combine locality preserving property with discriminant information to enhance the performance of LPP in pattern analysis. Being unsupervised, the original LPP does not make use of class membership relation of each point to be projected. To complement the original LPP, a supervised LPP has been proposed in the paper, called SLPP. The name implies that membership information influences on which points are included in the neighborhood of each point. That is to say, SLPP employs prior information about the original data set to perform learning the feature subspace.

The essence of SLPP algorithm is how to choose the similarity measure matrix \(S\) in Eq. (7). In LPP, \(S\) is only related to the neighborhood or the nearest neighbors. In other words, the selection of \(S\) is independent of class information though the samples in neighborhood may belong to the same class. Due to the locality preserving property of LPP, the points corresponding to samples would be close to each other in the reduced space if they are in neighborhood, though they may belong to different classes. This will result in an unfavorable situation in pattern analysis especially in classification problem.

Let us rearrange the order of samples in the original data set which will not affect the procedure of the algorithm. Suppose that the first \(M_1\) columns of \(X\) are occupied by the data of the first class, the next \(M_2\) columns are composed of the second class, etc., i.e. data of a certain class are compactly stored...
in \( X \). This step is of benefit to simplify the explanation of SLPP algorithm. As a consequence, \( X \) is changed to be an orderly matrix which is composed of submatrices \( A_i \) of size \( n \times M_i \), \( i = 1, \ldots, c \), where \( c \) is the number of classes. In the same manner, \( B_2, \ldots, B_c \) are generated by repeating the same process. Then the nearest neighbors for each \( x_j \in A_1 \) are sought in \( A_1 \) only. When applied to all \( x_j \in A_1 \), the procedure leads to a construction of the matrix \( B_1 \). When obtained \( B_1, \ldots, B_c \), the similarity measure matrix \( S \) is constructed by taking \( B_i \) as its diagonal structural elements, i.e.:

\[
S = \begin{bmatrix}
B_1 \\
B_2 \\
\vdots \\
B_c
\end{bmatrix}
\]  

(11)

To simplify the similarity computation between different points, we just set the weight equal to 1 if the two points belong to the same class. Therefore, \( B_i \) takes the following form:

\[
B_i = \begin{bmatrix}
0 & 1 & \cdots & 1 \\
1 & 0 & \cdots & 1 \\
\vdots & \ddots & \ddots & \vdots \\
1 & \cdots & 1 & 0
\end{bmatrix}_{M_i}
\]  

(12)

We can find that \( S \) is a symmetric matrix since each structural element \( B_i \) is symmetric.

The algorithmic procedure of SLPP is then stated as

1. **Order rearrangement**: As we have described in paragraph 4 of this section, samples of a certain class are stored compactly in original samples matrix \( X \) after order rearrangement.
2. **PCA projection**: For the sake of avoiding singularity of \( XDX^T \) and of reducing noise, we project \( X \) to its PCA subspace. \( X \) is still used to denote samples in PCA space, and the transformation matrix is denoted by \( W_{PCA} \).
3. **Computing similarity measure matrix \( S \)**: Because we have finished the order rearrangement of samples, \( B_i \) in (12) is easily computed. Then \( S \) is constructed by taking \( B_i \) as its diagonal structural elements.
4. **Eigenmap**: Solve the generalized eigenvector problem:

\[
XLX^T W = \lambda XDX^T W.
\]  

(13)

Then the final transformation matrix from original sample space to the embedded feature space is

\[
W_{SLPP} = W_{PCA} W,
\]  

(14)

where \( W \) is the solution of (13).

At a first glance, it is hard to evaluate the influence of the changed selection for similarity matrix \( S \) on the final feature space. Although the change is small, the influence is significant because \( S \) is used to calculate another matrix \( W \) which determines embedded coordinates. As a result, the reduced feature spaces obtained with unsupervised and supervised LPP are different. To visualize the differences, SLPP and LPP were applied to a small database from Simon Lucas, together with PCA and LDA. This database is composed of digits “0”–“9” and capital “A”–“Z”. The size of images in the database is \( 20 \times 16 \). We just took “A”–“Z” out of the database and randomly selected 13 samples for each class. Using PCA, LDA, LPP and SLPP algorithms, we projected the samples to the 2D embedded space. The visualization results are shown in Fig. 1.

To be convenient, the description of “V–Z” is absent from the right column of each figure though the corresponding samples are pictured using different color or symbols. When projected to 2D space using our SLPP algorithm, the samples belonging to the same class construct a nearly compact set. This is interesting and exciting! However, PCA, LDA and LPP projections do not demonstrate such excellent performances. Many samples are mixed up when projected to 2D space and it is difficult for us to distinguish them which class they belong to.

4. **Gabor wavelets**

The 2D Gabor functions proposed by Daugman are local spatial band-pass filters that achieve the theoretical limit for conjoint resolution of information in the 2D spatial and 2D Fourier domains, that is, Gabor wavelets exhibit desirable characteristics of spatial locally and orientation selectivity [21]. Donato et al. [22] had shown through experiments that the Gabor wavelet representation gives better performance than other techniques for classifying facial actions [27].
The Gabor wavelets (kernels, filters) can be defined as
\[
\Psi_{\alpha, \beta}(z) = \frac{\|k_{\alpha, \beta}\|^2}{\sigma^2} e^{-\frac{||k_{\alpha, \beta}||^2 + z^2}{2\sigma^2}} \left[ e^{ik_{\alpha, \beta}z} - e^{-\sigma^2/2} \right],
\]
where \(z = (x, y)\), and the wave vector \(k_{\alpha, \beta}\) is defined as
\[
k_{\alpha, \beta} = k_{\max} e^{i\phi_{\alpha}},
\]
where \(k_{\beta} = k_{\max}/f^\beta\) and \(\phi_{\alpha} = \pi\alpha/8\). \(k_{\max}\) is the maximum frequency, and \(f\) is the spacing factor between kernels in the frequency domain.

By scaling and rotation of \(k_{\alpha, \beta}\), all self-similar Gabor kernels in (15) can be generated from one filter, the mother wavelet. Each kernel is a product of a Gaussian envelope and a complex plane wave. In the square brackets in (15), the first term and the second term denote the oscillatory part and the DC part of the kernel, respectively. If the parameter \(\sigma\) has sufficiently large value, the effect of DC term becomes negligible. Here the parameters of \(\alpha\) and \(\beta\) are set to eight and five, respectively. In this paper, we also set \(\alpha\) and \(\beta\) to be eight and five. Fig. 2 shows the real part of the Gabor kernels at five scales and eight orientations and their magnitudes, with the following parameters: \(\alpha = \{0, 1, \ldots, 7\}\), \(\beta = 0, 1, \ldots, 4\), \(k_{\max} = \pi/2\) \(f = 2\), \(\phi = 2\pi\). The Gabor kernels show desirable performance of orientation selectivity, frequency and spatial locality.

Let \(f(x, y)\) be the gray level distribution of the image. The Gabor wavelet representation of the image is the convolution of \(f(x, y)\) with a series of Gabor kernels at different scales and orientation:
\[
Y_{\alpha, \beta}(z) = f(z) \ast \Psi_{\alpha, \beta}(z),
\]
(17)
where $\| \ast \|$ denotes the convolution operator, $Y_{x,\beta}$ is the corresponding convolution result related to different orientation $x$ and $\beta$. Applying the convolution theorem, it gives

$$Y_{x,\beta}(z) = \mathcal{F}^{-1}\{\mathcal{F}\{f(z)\}\mathcal{F}\{\Psi_{x,\beta}(z)\}\},$$

(18)

where $\mathcal{F}$ denote the Fourier transform.

Given a simple face image, Fig. 3 shows the Gabor representation: the real part and the magnitude. From Fig. 3, we can find out that these representations display desirable locality and orientation performance corresponding to the Gabor kernels in Fig. 2. In order to encompass all frequency and locality information as much as possible, this paper, same as Liu [23], concatenated the all Gabor representations at the five scales and eight orientations. Before the concatenation, $Y_{x,\beta}(z)$ is down-sampled by a factor $\rho$ to reduce the space dimension, and normalized to zero mean and unit variance. We then construct a vector out of the $Y_{x,\beta}(z)$ by concatenating its rows (or columns). Now let $Y^\rho_{x,\beta}(z)$ denote the normalized vector constructed from $Y_{x,\beta}(z)$, the augmented Gabor feature vector $Y^\rho$ is defined as

$$Y^\rho = (Y^\rho_{x,\beta} | x = 0, \ldots, 7; \beta = 0, \ldots, 4).$$

Then $Y^\rho$, which is a row vector or column vector, serves as the original space performed different methods for recognition.

5. Experimental results of GSLPP

The proposed GSLPP method applies SLPP to the augmented Gabor feature vector derived from the Gabor wavelet representation as described in (19). In this section, we applied the proposed GSLPP to the face recognition task, together with PCA, LDA, LPP, GPCA and GLDA. Before carrying out the experiments, each image will be transformed into its Gabor representation in (19) both in training process and in testing process. The face data sets include the well-known AR data set and CMU PIE data set. The details of the experiments are stated as follows.

5.1. AR face dataset

AR face data set consists of 26 frontal images with different facial expressions, illumination conditions, and occlusions for 126 subjects (70 men and 56 women) [24]. The images taken from two sessions of one specific person are shown in Fig. 4.

We select 80 individuals (50 men and 30 women) for our experiment. Thirteen images in each session were all chosen for every subject. That is, there are 26 images per subject in our database. Totally, our
database includes 2080 images. Then these images were the input of a face detection system which combines skin-based method and boosting method [25,26]. After face detection, the detected face images are converted to grayscale and resized to $40 \times 40$. In fact, there are still a very small number of faces not detected by the detection system although it achieves over 98% detection rate on our database. For the purpose of experiment, these undetected face images were cropped manually. The flow chart of detection process is shown in Fig. 5.

The face images wearing glasses and scarf in the first session and second session are contained in the training set and the testing set, respectively. The left 14 images are divided randomly into two parts: seven images for training and the other seven for testing. That is, there are 1040 images in the training set, and 13 images per subject. It is the same to the testing set. In terms of theoretical analysis, different

Fig. 4. Some samples of AR face data set.

Fig. 5. Face detection and preprocessing of AR.
methods (PCA, LDA, and LPP) would result in different embedding feature spaces. In fact, the basis of the feature space is the eigenvectors of the corresponding method. Furthermore, we can display these eigenvectors as images. Because these images look like human faces, they are often called Eigenfaces, Fisherfaces, Laplacianfaces [11,16,24]. For the eigenvectors of our SLPP, it may be called S-Laplacianfaces.

The S-Laplacianfaces derived from the training set are shown in Fig. 6, together with Eigenfaces, Fisherfaces and Laplacianfaces. The nearest neighbor classifier is employed for classification. The recognition rates reach the best results with 121, 101, 79, 79, 109, 82 dimensions for PCA, GPCA, LDA, GLDA, LPP and GSLPP, respectively. Fig. 7 shows the recognition rates vs. dimensionality reduction.

GSLPP method achieves the best recognition rate compared with the other five methods, though the recognition accuracy obtained by all these methods is relatively low. One possible reason is that there are some occluded images in training set and testing set. These images severely deteriorate the determinant power of these methods besides the exaggerated expression of non-occluded images. But the most important we want to demonstrate is that GSLPP is more effective than the other five methods. And the experiment proves its good performance.

5.2. CMU PIE face data set

The CMU PIE face data set consists of 68 subjects with 41,368 face images as a whole [27]. In this subsection, we selected 40 subjects in the database and used 140 face images for each individual. Seventy images are for training, and

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recognition Rate</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>49.1%</td>
<td>121</td>
</tr>
<tr>
<td>GPCA</td>
<td>54.5%</td>
<td>101</td>
</tr>
<tr>
<td>LDA</td>
<td>52.9%</td>
<td>79</td>
</tr>
<tr>
<td>GLDA</td>
<td>65.5%</td>
<td>79</td>
</tr>
<tr>
<td>LPP</td>
<td>50.2%</td>
<td>109</td>
</tr>
<tr>
<td>GSLPP</td>
<td>74.3%</td>
<td>85</td>
</tr>
</tbody>
</table>

Fig. 7. Experimental results on AR face data set: (a) the best recognition rates; (b) recognition rates vs. dimensions.

Fig. 8. Some samples of CMU PIE face data set.
the other Seventy images for testing. Still, faces are detected by the face detection system stated in Section 5.1. The detected faces are converted to grayscale and resized to 40 x 40, no other preprocessing. Some samples are shown in Fig. 8. Totally, there are 2800 images in the training set and the testing set, respectively.

For the sake of visualization, we illustrate S-Laplacianfaces derived from the training set, together with Eigenfaces, Fisherfaces and Laplacianfaces in Fig. 9. Still, nearest neighbor classifier was adopted to perform the recognition task for its simplicity, though there are other classifiers for pattern recognition such as Neural Network [28], Bayesian [29], Support Vector Machine [30], etc.

The recognition rates approach the best with 115, 95, 39, 39, 107 and 77 dimensions for PCA, GPCA, LDA, GLDA, LPP and GSLPP, respectively. Fig. 10 illustrates the recognition rates vs. dimensionality reduction.

The experiments above show that our GSLPP method achieves the best recognition performance with a relatively low dimension. Furthermore, it is robust to variations of illumination and facial expression. This can be illustrated by the comparative experiments on AR and PIE data sets. The superiorities of our GSLPP originate in the discriminating Gabor features on one hand, and the discriminating SLPP features on the other hand. The combination of these two properties makes GSLPP perfect in the above experiments compared with other five methods.

6. Discussion and future work

A general framework for Gabor-based supervised LPP was proposed in this paper. Experiments on a number of data sets demonstrated that GSLPP is a powerful feature extraction method, which when coupled with simple classifiers can yield promising recognition results. GSLPP takes the class membership information into account besides holding the locality preserving property of LPP and Gabor superior representation properties. It is noted that the supervision information Gabor property can enhance the performance of LPP respectively, to some extent. Since both local structure and discriminant information are important for classification, GSLPP outperforms the traditional LPP(also GLPP and SLPP), together with PCA, LDA, GPCA and GLDA which tend to preserve the global structure though LDA involves class information.
It is worthwhile mentioning that GSLPP takes advantage of more training samples, which is important to learn an efficient and effective embedding space representing the nonlinear structure of original patterns. Sometimes there might be only one training sample available for each class. In such a case, GSLPP cannot work since the similarity matrix is a nought matrix. How to overcome such problem is one of our future works.

In fact, it is not always the case that we can obtain all the class information in pattern analysis. Sometimes we only have unlabeled samples or partially labeled samples. Therefore, another extension of our work is to consider how to use these unlabeled samples for discovering the manifold structure and, hence, improving the classification performance [11,31,32]. We are now exploring these problems with full enthusiasm.

Acknowledgment

The authors wish to acknowledge that this work is supported by Fundamental Project of Science and Technology Department of Zhejiang Province under the Grant number 20060495.

References


