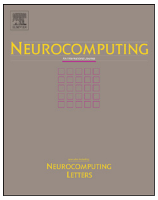
Neurocomputing 173 (2016) 290–297

Contents lists available at ScienceDirect

Neurocomputing

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



Graph Regularized Sparsity Discriminant Analysis for face recognition

Songjiang Lou [a,*], Xiaoming Zhao [a], Yuelong Chuang [a], Haitao Yu [b], Shiqing Zhang [a]

[a] *Institute of Image Processing & Pattern Recognition, Tai Zhou University, Taizhou, Zhejiang 318000, China*
[b] *College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China*

ARTICLE INFO

Article history:
Received 28 July 2014
Received in revised form
11 April 2015
Accepted 16 April 2015
Chennai Guest Editor
Available online 1 September 2015

Keywords:
Sparse representation
Graph embedding
Sparsity preserving projection
Feature extraction
Face recognition

ABSTRACT

Manifold learning and Sparse Representation Classifier are two popular techniques for face recognition. Because manifold learning can find low-dimensional representations for high-dimensional data, it is widely applied in computer vision and pattern recognition. Most of the manifold learning algorithms can be unified in the graph embedding framework, where the first step is to determine the adjacent graphs. Traditional methods use k nearest neighbor or the ε-ball schemes. However, they are parametric and sensitive to noises. Moreover, it is hard to determine the size of appropriate neighborhoods. To deal with these problems, in this paper, Graph Regularized Sparsity Discriminant Analysis, GRSDA, for short, is proposed. Based on graph embedding and sparsity preserving projection, the weight matrices for intrinsic and penalty graphs are obtained through sparse representation. GRSDA seeks a subspace in which samples in intra-classes are as compact as possible while samples in inter-classes are as separable as possible. Specifically, samples in the low-dimensional space can preserve the sparse locality relationship in the same class, while enhancing the separability for samples in different classes. Hence, GRSDA can achieve better performance. Extensive experiments were carried out on ORL, YALE-B and AR face databases, and the results confirmed that the proposed algorithm outperformed LPP, UDP, SPP and DSNPE.

1. Introduction

Face recognition, as one popular application of pattern recognition and computer vision, has aroused great interest among researchers. The main steps for face recognition include preprocessing, feature extraction and classification. In order to make the subsequent tasks easier, many algorithms have been proposed for the preprocessing part, like detection [1,2], and a survey on face detection can be found in [3]. Classification for face recognition has developed from the simple yet elegant nearest neighbor (NN) [4] method to recently proposed regression-based classification algorithms such as Linear Regression Classifier (LRC) [5], Sparse Representation Classifier (SRC) [6] and Collaborative Representation Classifier (CRC) [7]. These three regression-based algorithms have achieved comparable results and they have shown to have a great potential in practical applications. Moreover, in [8–10] many extensions to the aforementioned classification algorithms have also been proposed. Besides these, Support Vector Machine (SVM) [11] and deep learning [12,13] are also very popular in face recognition. For example, when deep learning was applied to face recognition, one algorithm is called DeepFace [12], which can achieve very impressive results in the Labeled Faces in the Wild (LFW) dataset.

Among all the algorithms for face recognition, appearance-based subspace learning schemes attract considerable interest due to its simplicity and desirable performance. Because the dimensionality of face images is usually very high, dimensionality reduction, which is also called feature extraction, is a key issue for face recognition, and has received tremendous attention in the past 20 years.

Many applications in computer vision and pattern recognition fields, such as face recognition, content-based image retrieval, bioinformatics etc., often confront high-dimensional and nonlinear samples. Nevertheless, dimensionality reduction gives an effective way to avoid the curse of dimensionality [14]. A lot of algorithms have been proposed in the past decades, and the two widely-used classic techniques are Principal Component Analysis (PCA) [15] and Linear Discriminant Analysis (LDA) [16,17], which are both matrix-decomposition based approaches [18] and assume that the distribution of samples is globally linear. However, in many applications such as face images where high-dimensional data are considered, distribution of samples is often nonlinear. One way to handle this problem is to use the kernel trick where the data from the original space is mapped to a higher-dimensional space. In the kernel space data is assumed to be linearly separable. Kernel Principal Component Analysis (KPCA) [19] and Kernel Linear Discriminant Analysis (KLDA) [20] are two representatives and can find their effectiveness in pattern recognition. However, how to choose the appropriate kernel is not an easy task, as it often influences the success of the algorithms.

[*] Corresponding author. Tel.: +8613586117695.
E-mail address: lousongjiangac@163.com (S. Lou).

http://dx.doi.org/10.1016/j.neucom.2015.04.116
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Another category is manifold learning algorithms (for instance, ISOMAP [21], Locally Linear Embedding (LLE) [22], Laplacian Embedding (LE) [23], Local Tangent Space Alignment (LTSA) [24], Parallel vector Field Embedding (PFE) [25], Geodesic Distance Function Learning (GDL) [26], and Parallel Field Alignment for cross media Retrieval (PFAR) [27]), which have been proposed to discover the intrinsic low-dimensional presentations for high-dimensional and nonlinear data. However, these kinds of algorithms cannot map a new sample to the corresponding low-dimensional space, which is also called the Out-Of-Sample extension problem [28]. Thus numerous methods have been proposed to solve this problem, which can achieve an explicit mapping, like Locality Preserving Projection (LPP) [29,30], Neighborhood Preserving Embedding (NPE) [31], Unsupervised Discriminant Projection (UDP) [32], Marginal Fisher Analysis (MFA) [33], Linear Discriminant Embedding (LDE) [34], Orthogonal LPP [35], locality preserving discriminant projections (LPDP) [36], Discriminative multi-manifold analysis [37] and Iterative Subspace Analysis based on Feature Line Distance [38]. These algorithms considered the manifold structure or the discriminant information in one way or another, and showed to be more efficient than the traditional methods in some special scenarios. For example, in [37] the inter-manifold and intra-manifold graphs were defined according to the label information and then the optimal projection was searched, which yielded impressive results for a special case where only one sample was available for each person.

All these methods can be unified in the graph embedding framework [39,40]. In this framework the first step is to construct the graphs, that is, the intrinsic and penalty graphs. However, the performance of the algorithms is heavily dependent on how to construct the graphs. Traditional schemes employ k nearest neighbor or the ε-ball method, but how to choose the appropriate neighborhood size or the ball radius remains unclear. Moreover, for these two methods, the graph construction and the weight assignment are independent. One ideal model is that there is no parameter, and the graph construction and the weight assignment can be finished in one step [41].

Sparse representation [6,42] has received considerable interest in the last few years. The main idea in it is that the given test sample can be represented as a linear combination of the training samples, and the classification is achieved by evaluating which class leads to the minimum reconstructive deviation. The coefficients obtained by sparse representation can reflect the contributions of the samples to reconstruct the given test sample. It is reported in [43] that the bigger coefficients were the more likely these samples belonged to the same class. Hence, the reconstruction coefficients can be considered as a measurement of similarity.

Motivated by this idea, some researchers attempted to construct the adjacent graphs in a nonparametric way in which the graph construction and weight assignment can be finished in one step, and this technique has been applied to a wide range of applications due to the fact that it was parameter-free and robust to noises.

Yan et al. [44,45] proposed l_1 graph for image analysis, and constructed the graphs by sparse representation. In [46] a graph regularized sparse coding method was proposed, which combined local manifold structure into sparse representation. However, it was unsupervised and the performance was limited to some extent. Similar to l_1 graph, Qiao et al. [47] proposed Sparsity Preserving Projection (SPP), in which every sample was presented as a linear combination of the remaining samples. SPP tried to find a projection which can preserve the sparse reconstructive relationship. There was no need to choose the parameter of neighborhood size, and the authors pointed out that it had natural discriminative power and was robust to noises to some extent. However, SPP took the whole training set as the dictionary, and it was an unsupervised method. Zhang et al. [48] introduced a graph optimization for dimensionality reduction with sparsity constraints (GODRSC) which attempted to learn the sparse representation coefficients and the embedding simultaneously. In [49] Sparse Representation Classifier Steered Discriminant Projection (SRCS-DP) was proposed, which tried to find a projection by maximizing the inter-class reconstruction error while minimizing the intra-class reconstruction error. Therefore it had more discriminative power, but it neglected the manifold structure and was time-consuming due to the fact that the projection matrix and sparse presentation coefficients were obtained iteratively. Chen and Jin [50] proposed a new feature extraction method called Reconstructive Discriminant Analysis (RDA) from the viewpoint of linear regression classification. Gui et al. [51] designed a new scheme called Discriminant Sparse Neighborhood Preserving Embedding (DSNPE), which represented the data as a linear combination of samples from the same class and preserved the sparse reconstructive relationship in the same class. However, it ignored the inherent manifold structure of training samples, especially the inter-class manifold structure, as it only integrated SPP and maximum margin criterion (MMC) [52]. Similar works can also be found in [53,54]

To exploit the merits of manifold learning and robustness of sparse representation, this paper presents a new algorithm called Graph Regularized Sparsity Discriminant Analysis (GRSDA), which utilizes sparse representation as a way to graph construction and weight assignment. In GRSDA, the intrinsic and penalty graphs are constructed via sparse representation and the weights are obtained subsequently, so it avoids the difficulty of determining the neighborhood size. On the one hand, it inherits the property of preserving the manifold structure like LPP; on the other hand, it derives from LDA which has good discriminative power. Under the graph embedding framework, GRSDA seeks a subspace, where samples from the same class are as compact as possible, while samples from different classes are as separable as possible.

The rest of this paper is organized as follows: Section 2 presents an overview of the related works like sparse representation, sparsity preserving projection and graph embedding. Graph Regularized Sparsity Discriminant Analysis is proposed in Section 3. Experiment results for the proposed algorithm and the related algorithms are shown in Section 4. Section 5 gives the conclusion.

2. The related work

Suppose that we have a training set $X = \{X_1, X_2, \ldots, X_C\} = \{x_1, x_2, \ldots, x_N\}$ of n samples, where $x_i \in R^D$ $(i = 1, 2, \ldots, N)$ and D is the dimensionality. There are C classes, and there are N_k $(k = 1, 2, \ldots, C)$ samples in the kth class. The aim of dimensionality reduction is to seek a projection A, so that every sample in the original space can be mapped to a low-dimensional space by $y_i = A^T x_i \in R^d$, where $d \ll D$.

2.1. Sparse representation

If a given test sample y belongs to the ith class, sparse representation assumes that y can be represented as a linear combination of the training samples in the ith class $X_i = \{x_i^1, x_i^2, \cdots, x_i^{N_i}\}$. In other words, we can present y as follows:

$$y = w_i^1 x_i^1 + w_i^2 x_i^2 + \cdots + w_i^{N_i} x_i^{N_i} = X_i W_i \tag{1}$$

where W_i denotes the representation coefficient of y over X_i. Ideally, the representation coefficients of other classes are zero, that is, $W_j = 0, \forall j \neq i$. Thus y can be represented as a linear combination of the whole training samples if we stack X_i, $i = 1, 2, \cdots, C$ into the whole training set X. Therefore we can obtain the representation coefficient matrix $W = [W_1, W_2, \ldots, W_C]$. In other words:

$$y = 0 + 0 + \cdots, 0 + w_i^1 x_i^1 + w_i^2 x_i^2 + \cdots + w_i^{N_i} x_i^{N_i} + 0 + 0 + \cdots + 0 \\ = XW \tag{2}$$

The above model can be expressed as follows:

$$\tilde{W} = \arg\min \ \|W\|_0, \quad s.t. \quad y = XW \tag{3}$$

where $\|W\|_0$ is the l_0-norm, which denotes the number of non-zero entries in the vector W.

The above optimization problem cannot be solved in polynomial time since the l_0-norm optimization problem is an NP problem. Fortunately, recent efforts on Compressed Sensing [55] have been made and demonstrated that the l_0-norm is equivalent to the l_1-norm optimization problem if the solution is sparse enough [56], which can be solved by the following optimization problem, that is:

$$\tilde{W} = \arg\min \ \|W\|_1, \quad s.t. \quad y = XW \tag{4}$$

After obtaining the optimal sparse representation $\tilde{W} = [\tilde{W}_1, \tilde{W}_2, \cdots, \tilde{W}_C]$, the reconstructive errors between the test sample and the samples in the ith class are calculated:

$$r_i(y) = \|y - X_i \tilde{W}_i\|, \ i = 1, 2, \cdots, C \tag{5}$$

Then the label of y is determined as the class with the minimal reconstructive error:

$$label(y) = arg \min_{i=1,2,\cdots,C} r_i(y) \tag{6}$$

2.2. Sparsity preserving projection

Sparsity preserving projection aims to preserve the sparse reconstruction relationship, and its objective function is:

$$\min_A \sum_{i=1}^{N} \|A^T x_i - A^T X w_i\| \\ s.t. \quad A^T X X^T A = I \tag{7}$$

Where A is the transformation matrix and I is an identity matrix. Furthermore, w_i is the sparse reconstruction coefficients over X for x_i, and it can be obtained by the following optimization problem:

$$\min_w \|w_i\|_0 \quad s.t. \quad x_i = X w_i \tag{8}$$

Where $\|w_i\|_0$ denotes the number of non-zero entries in w_i, and the set X represents the whole training set excluding x_i. Thus, we have $X = \{x_1, x_2, .., x_{i-1}, 0, x_{i+1}, \ldots, x_N\}$. However, the above optimization problem is non-convex. The authors in [56] pointed out that l_0-norm can be approximately solved by the l_1-norm if w_i is sparse enough, so (8) is approximately equivalent to the following optimization problem, which can be efficiently solved by linear programming

$$\min_w \|w_i\|_1 \ s.t. \quad x_i = X w_i \tag{9}$$

If we compute the weight vector w_i for every x_i, $i = 1, 2, \cdots, N$, we can get the sparse reconstructive weight matrix as $W = [w_1, w_2, \ldots, w_N]^T$. More details can be found in [47].

By simple algebra (7) can be reduced to the following problem:

$$\min \ tr(A^T X(I - W - W^T + W^T W)X^T A) \\ s.t. \quad A^T X X^T A = I \tag{10}$$

2.3. Graph embedding

Graph embedding tries to preserve the local neighboring relationship in the low-dimensional space. The samples and their relationships are denoted as a graph $G = (X, W)$, where vertices X are composed of $x_i(i = 1, 2\ldots, N)$, W is the adjacent matrix, and its entry W_{ij} denotes the similarity between x_i and x_j. Its objective function is:

$$\min_A J(A) = \sum_{i=1}^{N}\sum_{j=1}^{N} \|A^T x_i - A^T x_j\|^2 W_{ij} \\ s.t. \ A^T X D X^T A = I \tag{11}$$

By simple algebra (11) can be reduced to the following problem:

$$\max_A J(A) = \frac{tr(A^T X D X^T A)}{tr(A^T X L X^T A)} \tag{12}$$

Where D is a diagonal matrix, $D_{ii} = \sum_j W_{ij}$, and $L = D - W$ is a Laplacian matrix.

3. Graph Regularized Sparsity Discriminant Analysis

3.1. The proposed algorithm

Graph embedding is a representative manifold learning algorithm, and LDA, LPP, UPD and MFA can be unified in the graph embedding framework with respect to different L and Dmatrices. The first step of graph embedding is graph construction, and the mostly adopted methods are k nearest neighbor and the ε-ball method. However, these two methods are parametric and the performance of the algorithms is highly dependent on the choice of the parameter value. In [47] it was reported that the adjacent graph structure and graph weights are highly interrelated and should not be separated. Hence, it is desirable to design a method that can simultaneously carry out these two tasks in one step. In this section, we propose a new method called Graph Regularized Sparsity Discriminant Analysis, which incorporates the intra-class and inter-class discriminant information into the graph construction by sparse representation, and then calculate the intra-class adjacent matrix $W$and the inter-class adjacent matrix B.

In the intra-class graph, the weight w_{ij} can reflect the similarity between x_i and x_j. The bigger the w_{ij}, the more likely the x_i and x_j from the same class. In sparse representation, every sample can be represented as a linear combination of the remaining samples, and the coefficient w_{ij} means the contribution of x_j to reconstruct x_i. And it is reported in [6] that samples can be effectively reconstructed by the samples from the same class. Hence, x_i^j can be represented as a linear combination of samples from the same class:

$$\min \|w_i^j\|_1 \\ s.t. \ x_i^j = X_i w_i^j \tag{13}$$

In (13) x_i^j denotes the jth sample in the ith class and X_i denotes the samples from the ith class except x_i^j. We can get the reconstruction coefficients for x_i^j through Eq. (13), and it can be written as a vector $w_i^j = (\omega_i^1, \omega_i^2, \ldots, \omega_i^{j-1}, 0, \omega_i^{j+1}, \ldots, \omega_i^{N_i})^T$. By the same way, we can get the sparse representation coefficients for the i^{th} class in the form of $W_i = [w_i^1, w_i^2, \cdots, w_i^{N_i}]$ Moreover, for all the samples, we can get the intra-class weight matrix as

$$W = \begin{bmatrix} W_1 & O & \cdots & O \\ O & W_2 & \cdots & O \\ \vdots & \ddots & & \vdots \\ O & \cdots & & W_C \end{bmatrix}$$

Similar to graph embedding, we hope that in the low-dimensional space, samples from the same class are as compact as possible. Therefore we construct the intrinsic graph $G_{intrinsic} = (X, W)$, and the objective function is:

$$\min_A J_1(A) = \sum_{i=1}^{N} \sum_{j=1}^{N} \|A^T x_i - A^T x_j\|^2 W_{ij} \tag{14}$$

It can be simplified as:

$$\min_A J_1(A) = \sum_{i=1}^{N} \sum_{j=1}^{N} \|A^T x_i - A^T x_j\|^2 W_{ij} \\ = 2\, tr(A^T X L_w X^T A) \tag{15}$$

Where $L_w = D_w - W$is the Laplacian matrix, D_w is a diagonal matrix and $D_{w\ ii} = \sum_j W_{ij}$.

On the other hand, every sample can be represented as a linear combination of samples from different classes, that is

$$\min \|b_i^j\|_1 \\ s.\,t.\ \ x_i^j = X^i b_i^j \tag{16}$$

In (16) x_i^j denotes the jth sample in the ith class and X^i denotes samples except the ones from the ith class, that is $X^i = (X_1, \ldots, X_{i-1}, O, X_{i+1}, \ldots, X_C)$.
We can get the optimal sparse reconstruction vector as $\tilde{b}_i^j = (\tilde{b}_i^{j,1}, \tilde{b}_i^{j,2}, \ldots, \tilde{b}_i^{j,(N_1+N_2+\ldots+N_{i-1})}, \overbrace{0, 0, \ldots, 0}^{N_i}, \tilde{b}_i^{j,(N_1+N_2+\ldots+N_i+1)}.$
$\tilde{b}_i^{j,(N_1+N_2+\ldots+N_i+2)}, \ldots, \tilde{b}_i^{j,N})^T$

According to the label information, we can rewritten $\tilde{b}_i^j$ in the class-wise form, $\tilde{b}_i^j = (\beta_i^{j,1}, \beta_i^{j,2} \ldots, \beta_i^{j,i-1}, O, \beta_i^{j,i+1}, \ldots, \beta_i^{j,C})$. Similar to the intra-class weight matrix, we can get the inter-class weight vector for the ith class in the form of$\tilde{B}_i = [\tilde{B}_i^1, \tilde{B}_i^2, \cdots \tilde{B}_i^{i-1}, O, \tilde{B}_i^{i+1}, \cdots, \tilde{B}_i^C]$, where $\tilde{B}_i^k = [\beta_i^{1,k}, \beta_i^{2,k} \ldots, \beta_i^{N_i,k}]^T$denotes the weight matrix of the kth class to reconstruct the ith class. Subsequently we can get the inter-class weight matrix for the whole training set as

$$\tilde{B} = \begin{bmatrix} O, & \tilde{B}_1^2, & \tilde{B}_1^3 \cdots & \tilde{B}_1^C \\ \tilde{B}_2^1, & O, & \tilde{B}_2^3 & \cdots \tilde{B}_2^C \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{B}_C^1, & \tilde{B}_C^2, & \tilde{B}_C^3 & \cdots O \end{bmatrix},$$

where the entry $\tilde{B}_{ij}$ in $\tilde{B}$denotes the contribution of x_j to reconstruct x_i. The bigger the $\tilde{B}_{ij}$, the more similar the x_i and x_j. However, they are from different classes. A dissimilarity matrix B is defined, whose (i, j)thentry is defined by $B_{ij} = 1 - \tilde{B}_{ij}$. We hope that in the low-dimensional space, samples from different classes are as separable as possible. We construct the penalty graph$G_{penalty} = (X, B)$, and the objective function is:

$$\max_A J_2(A) = \sum_{i=1}^{N} \sum_{j=1}^{N} \|A^T x_i - A^T x_j\|^2 B_{ij} \tag{17}$$

Or in the simplified form:

$$\max_A J_2(A) = \sum_{i=1}^{N} \sum_{j=1}^{N} \|A^T x_i - A^T x_j\|^2 B_{ij} \\ = 2\, tr(A^T X L_b X^T A) \tag{18}$$

Where $L_b = D_b - B$ is the Laplacian matrix, and D_b is a diagonal matrix having entries $D_{bii} = \sum_j B_{ij}$.

Similar to Marginal Fisher Analysis, by combining (15) and (18), we get the following optimization problem:

$$\max J(W) = \frac{J_2(A)}{J_1(A)} \\ = \frac{tr(A^T X L_b X^T A)}{tr(A^T X L_w X^T A)} \tag{19}$$

The above optimization problem can be solved by the following generalized eigen-problem:

$$X L_b X^T a = \lambda X L_w X^T a \tag{20}$$

$S_b = X L_b X^T$is called the inter-class sparse scatter matrix, and $S_w = X L_w X^T$ is called the intra-class sparse scatter matrix. Suppose that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ are the d largest eigenvalues of (20) and its corresponding eigenvectors are $a_1, a_2, \ldots, a_d$. As a result, we get the projection matrix $A = [a_1, a_2, \ldots a_d]$. Hence, any sample x_i in the original space can be mapped to the low dimensional space by $y_i = A^T x_i$.

In applications, such as face recognition where the dimensionality is much larger than the number of faces, the intra-class sparse scatter matrix is usually singular. To make the intra-class sparse scatter matrix nonsingular, Principal Component Analysis (PCA) is implemented and 98% of the energy of the images is retained.

The algorithmic procedure of GRSDA can be formally summarized as follows:

Input: Training set $X = \{x_1, x_2, \ldots, x_N\} = \{X_1, X_2, \cdots, X_C\}, x_i \in R^D (i = 1, 2, \ldots, N)$, and its labels $l = (1, 2, \ldots, C)$. The size of the kth class is N_k and $N = \sum_{k=1}^{C} N_k$, whereas intrinsic dimensionality is d.

Output: projection matrix A.

Step1: For X, PCA is implemented as a preprocessing step and 98% of the energy of images is retained. For simplification, we still denote X as the training set after PCA.

Step2: For every sample in X, compute its sparse representations by (13) and (16). And construct its intrinsic graph $G_{intrinsic} = (X, W)$ and penalty graph $G_{penalty} = (X, B)$. For the sake of symmetry, we obtain W and B by $W = (W + W^T)/2$, $B = (B + B^T)/2$.

Step3: Calculate the intra-class sparse scatter matrix $S_w = X L_w X^T$ and inter-class sparse scatter matrix $S_b = X L_b X^T$.

Step4: Solve the generalized eigen-problem in (20) and get the d largest eigen-values $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ and the corresponding eigen-vectors $a_1, a_2, \ldots, a_d$. As a result, the projection matrix $A = [a_1, a_2, \ldots a_d]$ is obtained.

Step5: For every sample x_i in the original space, its low-dimensional representation is $y_i = A^T x_i$.

3.2. Connections to LDA and LPP

LPP attempts to preserve the local structure of samples. In other words, if two samples are close in the original space, they are also close to each other in the low-dimensional space. However, the performance of LPP is heavily dependent on the adjacent weights. Different graphs will lead to different variants of LPP. The proposed algorithm is an extension of LPP, where the adjacent weights for the intrinsic and penalty graphs are calculated via sparse representation. Unlike the conventional LPP which uses the k nearest neighbor or the ε-ball method to determine the neighborhood size, the proposed method is parameter-free. GRSDA can be considered as a regularized version of LPP by a modified l_1 regularization problem which encodes prior knowledge.

While LDA attempts to find the axis which maximizes the inter-class scatter and minimizes the intra-class scatter simultaneously, the proposed algorithm tries to find the axis that the sparse scatter for inter-class is maximized and the sparse scatter for intra-class is minimized. So clearly, the difference between LDA and GRSDA is the definition of scatter matrices.

Actually, LDA and LPP can be unified in the graph embedding framework. GRSDA is a variant of LDA and LPP. To compare them in terms of model parameters and the way to construct the graph, Table 1 shows the differences between them.

4. Experiments and analysis

To evaluate the effectiveness and the correctness of the proposed algorithm, experiments are carried out on ORL [57], extended YALE-B [58] and AR [59] databases, and the results are compared with LPP, UDP, SPP and DSNPE.

LPP, UDP, SPP, DSNPE and GRSDA are adopted to find the low-dimensional representations, and the nearest neighbor classifier is used for classification in the projected spaces. To make some matrices nonsingular, PCA is employed as a preprocessing step and 98% of the energy of images is retained. For LPP and UDP, the neighborhood size needs to be determined. In this paper the neighborhood size was set to be $k = n_i - 1$, where n_i is the number of samples in the ith class.

Table 1
Graph construction comparisons for LDA, LPP and GRSDA.

	Model parameters	Graph construction manners
LDA	No parameters	Globally
LPP	k or ε	Local neighborhood distance
GRSDA	No parameters	Intra-class and inter-class sparse representation

4.1. Experimental results on ORL database

The ORL database contains 40 individuals; each has 10 images with some variations in poses, facial expressions and some details. Each image was taken at different times and has different variations including expressions like open or closed eyes, smiling or non-smiling. Some of the images were captured with a tolerance for some tilting and rotation of the face up to 20°. Fig. 1 shows some samples from ORL database.

For ORL database, each image is resized to 32 × 32. We first randomly chose 4 images per subject to form a training set, and the remaining images of the database were taken to form a testing set. We repeated 20 times for each algorithm and used the best recognition rate among these 20 times as the recognition rate. Table 2 shows the best recognition rate (%) for each algorithm with its corresponding dimensionality.

To evaluate the effect of dimensionality on recognition rate, Fig. 2 shows the recognition rates for different methods with respect to different dimensionality.

4.2. Experimental results on extend YALE-B database

The extended YALE-B database includes 38 individuals, and each individual has 64 images which were captured under different poses and illuminations. Fig. 3 shows some samples from YALE-B database.

In our experiments, for the sake of efficient computation, each image is resized to 32 × 32. We first randomly chose 10, 20 and 30 images per subject to form a training set, and the remaining images of the database were taken to form a testing set. We repeated 20 times for each algorithm and used the best recognition rate among these 20 times as the recognition rate. Table 3 shows the best recognition rate (%) for each algorithm and the corresponding dimensionality is given in the parenthesis.

Table 2
Best recognition rates (%) with respect to the dimensionality for different methods on ORL database.

Methods	Recognition rates (%)	Dimensionality
LPP	90.4	39
UDP	91.6	33
SPP	92.9	63
DSNPE	93.5	68
GRSDA	93.8	58

Fig. 1. Samples from ORL.

4.3. Experimental results on AR database

The AR database contains 126 subjects (70 men and 56 women). There are more than 4000 color face images, including fontal views of faces, which were taken in two sessions under various conditions such as different facial expression, lighting conditions and occlusion (sunglasses and scarf). In our experiments, we chose a subset of 100 subjects (50 men and 50 women), and each had 14 different images. Hence, the total number of face images was 1400. The original image size was 165 × 120, but for efficient computation, we manually resized images to 50×40. Fig. 4 shows some samples from AR database.

In our experiments, We first randomly chose $l(l=3:4:5:6)$ images per subject to form a training set, and the remaining images of the database were taken to form a testing set. For each l, we repeated 20 times for each algorithm and used the best recognition rate among these 20 times as the recognition rate. Table 4 shows the best recognition rate (%) for each algorithm and the corresponding dimensionality is given in the parenthesis.

4.4. Computation time

To evaluate the computation efficiency, we present the computation time of different methods. The computation time is represented by the whole operation time for dimensionality reduction and classification. Our experiment environment is Intel CPU 2.10 GHz, 1G RAM memory, MATLAB 7.0.1 (R14). Table 5 shows the computation time of all used methods on the ORL database for 4 training sample per individual.

From the results in Table 5, we can observe that LPP and UPD are much faster than the sparse-representation-based feature extraction algorithm (SPP, DSNPE and GRSDA). This is probably because SPP, DSNPE and GRSDA involve the l_1 norm optimization problem, which is very time-consuming.

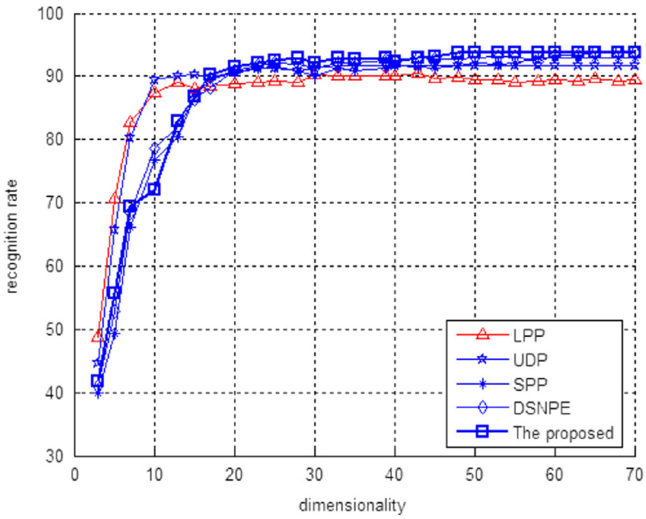


Fig. 2. Recognition rates vs. dimensionality on ORL database.

4.5. Discussions

From the experiments we can see that all algorithms perform better with the increased training number and it might be the fact that if there are more training samples, the features extracted can better represent the sample. We can also see that all algorithms perform better on ORL than on other databases. This is probably because the images on ORL have less variation than the images on YALE-B and AR. UDP outperforms LPP, and this is probably because LPP only considers the local structure, while UDP fully utilizes the local and non-local structure of data. SPP is based upon sparse representation, which preserves the sparse reconstructive relationship of the data. As the reconstruction coefficients can be seen as a measurement of similarity, that is to say, if the coefficient of one sample to reconstruct another is larger, they are more likely from the same class. So after projection, the samples from the same class are clustered to some extend, and the features extracted contain natural discriminant information even if it is unsupervised. DSNPE can preserve the sparse reconstructive relationship in the same class and maximize the margin between different classes, so it has better performance than SPP. The proposed algorithm, on the one hand, preserves the intra-class sparse locality relationships like LPP, on the other hand, maximizes the inter-class sparse scatter. So after projection, samples from the same class are compact while samples from different classes are far apart. Overall, the proposed algorithm has better recognition performance than LPP, UDP, SPP and DSNPE.

5. Conclusion

In this paper, a new algorithm for face recognition named Graph Regularized Sparsity Discriminant Analysis is proposed. Based on sparsity preserving projection, the weights for intrinsic and penalty graphs are obtained by sparse representation. This scheme avoids the difficulty of choosing the right parameter for manifold learning algorithm, and the graph construction and weight calculation are finished within one step. Then, by maximizing the inter-class sparse scatter while minimizing the intra-class sparse scatter, in the low-dimensional space, samples from

Table 3
Best recognition rates(%) with respect to the dimensionality for different methods on YALE-B database.

Number of training samples	LPP	UDP	SPP	DSNPE	GRSDA
10	68.5(218)	63.3(176)	76.4(198)	81.6(220)	82.7(266)
20	82.4(354)	81.8(298)	85.8(357)	87.6(340)	89.7(324)
30	86.2(478)	87.8(376)	90.2(421)	91.7(350)	93.4(361)

Fig. 3. Samples from YALE-B.

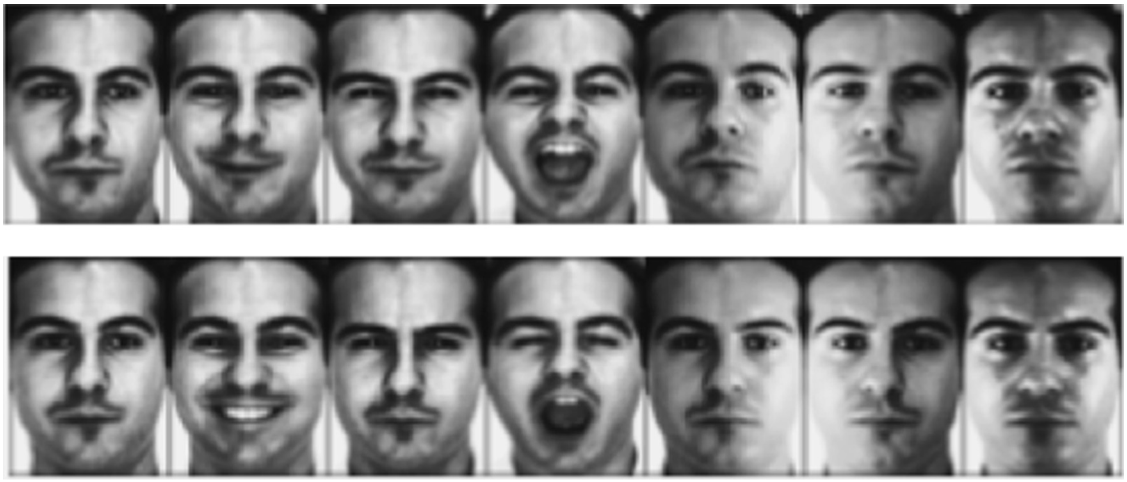


Fig. 4. Samples from AR database.

Table 4
Best recognition rates (%) with respect to the dimensionality for different methods on AR database.

Methods	Training number			
	3	4	5	6
LPP	75.2(125)	84.2(125)	87.9(110)	88.7(110)
UDP	76.3(115)	84.9(115)	88.2(110)	90.1(110)
SPP	72.4(150)	81.7(150)	89.5(145)	91.4(145)
DSNPE	73.8(120)	80.9(120)	89.8(135)	92.2(135)
GRSDA	74.2(130)	82.4(130)	90.7(130)	93.8(130)

Table 5
Comparisons of computation time in second on ORL database.

LPP	UDP	SPP	DSNPE	GRSDA
0.87	0.98	281.52	324.75	348.26

the same class are compact, while samples from different classes are far apart. Lastly, experiments are carried out on face recognition, and the results confirm that the proposed algorithm has superior performance compared with related algorithms. Moreover, the proposed method shows that its performance is not promising when the training numbers are small. We will study this problem in our future work.

Acknowledgments

The authors would like to thank all the anonymous reviewers for their valuable and constructive remarks that are very helpful to improve the quality of this paper. This work is supported by the Zhejiang Provincial Natural Science Foundation of China (No LQ15F020001) and Zhejiang Provincial Natural Science Foundation of China (No LQ15F020001) and the National Natural Science Foundation of China (Grant nos. 61272261 and 61203257).

References

[1] Y. Pang, H. Yan, Y. Yuan, K. Wang, Robust CoHOG feature extraction in human-centered image/video management system, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (2) (2012) 458–468.

[2] X. Jiang, Y. Pang, J. Pan, X. Li, Flexible sliding windows with adaptive pixel strides, Signal Process. 110 (2015) 37–45.

[3] E. Hjelmås, B.K. Low, Face detection: a survey, Comput. Vis. Image Underst. 83 (3) (2001) 236–274.

[4] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27.

[5] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2106–2112.

[6] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.

[7] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: which helps face recognition?, in: Proceedings of ICCV, 2011, pp. 471–478.

[8] Y. Xu, D. Zhang, J. Yang, J.-Y. Yang, A two-phase test sample sparse representation method for use with face recognition, IEEE Trans. Circuits Syst. Video Technol. 21 (9) (2011) 255–1262.

[9] Allen Y. Yang, et al., Fast-minimization algorithms for robust face recognition, IEEE Trans. Image Process. 22 (8) (2013) 3234–3246.

[10] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, Y. Ma, Towards a practical face recognition system: robust alignment and illumination by sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2) (2012) 372–386.

[11] Y. Pang, K. Zhang, Y. Yuan, K. Wang, Distributed object detection with linear SVMs, IEEE Trans. Cybern. 44 (11) (2014) 2122–2133.

[12] Y. Taigman, Ming Yang, M. Ranzato, L. Wolf, DeepFace: closing the gap to human-level performance in face verification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1701–1708.

[13] Y. Sun, Y. Chen, X. Wang, Deep learning face representation by joint identification–verification, Adv. Neural Inf. Process. Syst. (2014) 1988–1996.

[14] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[15] H. Abdi., L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev.: Comput. Stat. 2 (4) (2010) 433–459.

[16] A. Martinez, A. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233.

[17] Y. Pang, S. Wang, Y. Yuan, Learning regularized LDA by clustering, IEEE Trans. Neural Netw. Learn. Syst. 25 (12) (2014) 2191–2201.

[18] X. Li, Y. Pang, Deterministic column-based matrix decomposition, IEEE Trans. Knowl. Data Eng. 22 (1) (2010) 145–149.

[19] Bernhard Scholkopf, Alexander Smola, Klaus-Robert Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.

[20] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, Fisher discriminant analysis with kernels, in: Proceedings of the IEEE Signal Processing Society Workshop on Neural Network for Signal Processing, 1999, pp. 41–48.

[21] J.B. Tenenbaum, V. Silva, J.C. de, Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[22] S. Rowies, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[23] M. Belkin, P. Niyogo, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.

[24] Zhenyue Zhang, Hongyuan Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, SIAM J. Sci. Comput. 26 (1) (2004) 313–338.

[25] B. Lin, X. He, C. Zhang, Ming Ji, Parallel vector field embedding, J. Mach. Learn. Res. 14 (1) (2013) 2945–2977.

[26] B. Lin, J. Yang, X. He, J. Ye, Geodesic distance function learning via heat flow on vector fields, in: Proceedings of ICML, 2014, pp. 145–153.

[27] X. Mao, B. Lin, D. Cai, X. He, J. Pei, Parallel field alignment for cross media retrieval, in: Proceedings of ACM Multimedia, 2013, pp. 897–906.

[28] F. Dornaika, B. Raducanu, Out-of-Sample embedding for manifold learning applied to face recognition, in: Proceedings of CVPR, 2013, pp. 862–868.

[29] X. He, P. Niyogi, J. Han, Face recognition using Laplacian faces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[30] Y. Xu, A. Zhong, J. Yang, D. Zhang, LPP solution schemes for use with face recognition, Pattern Recognit. 43 (2) (2010) 4165–4176.
[31] X. He, D. Cai, S. Yan, Neighborhood preserving embedding, in: Proceedings of ICCV, 2005, pp. 1208–1213.
[32] J. Yang, D. Zhang, J.Y. Yang, et al., Globally maximizing, locally minimizing: unsupervised discriminant projection with application to face and palm biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007) 650–664.
[33] D. Xu, S. Yan, D. Tao, S. Lin, H.J. Zhang, Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval, IEEE Trans. Image Process. 16 (11) (2007) 2811–2821.
[34] Hwann-Tzong Chen, Huang-Wei Chang, Tyng-Luh Liu, Local discriminant embedding and its variants, in: Proceeding of CVPR, 2005, pp. 846–853.
[35] Deng Cai, Xiaofei He, Jiawei Han, Hong-Jiang Zhang, Orthogonal Laplacianfaces for face recognition, IEEE Trans. Image Process. 15 (11) (2006) 3608–3614.
[36] Jie Gui, Wei Jia, Ling Zhu, Shu-Ling Wang, De-Shuang Huang, Locality preserving discriminant projections for face and palmprint recognition, Neurocomputing 73 (13–15) (2010) 2696–2707.
[37] J. Lu, Y.P. Tan, G. Wang, Discriminative multimanifold analysis for face recognition from a single training sample per person, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 39–51.
[38] Y. Pang, Y. Yuan, X. Li, Iterative subspace analysis based on feature line distance, IEEE Trans. Image Process. 18 (4) (2009) 903–907.
[39] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, Stephen Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[40] Y. Pang, Z. Ji, P. Jing, X. Li, Ranking graph embedding for learning to rerank, IEEE Trans. Neural Netw. Learn. Syst. 24 (8) (2013) 1292–1303.
[41] Bogdan Raducanua, Fadi Dornaikab, Embedding new observations via sparse-coding for non-linear manifold learning, Pattern Recognit. 47 (1) (2014) 480–492.
[42] Weizhong Zhang, Lijun Zhang, Yao Hu, Rong Jin, Deng Cai, Xiaofei He, Sparse learning for stochastic composite optimization, in: Proceedings of AAAI, 2014, pp. 893–900.
[43] S. Wu, Spectral clustering of high-dimensional data exploiting sparse representation vectors, Neurocomputing 135 (2014) 229–239.
[44] Bin Xu, Jianchao Yang, Shuicheng Yan, et al., Learning with L1-graph for image analysis, IEEE Trans. Image Process. 19 (4) (2010) 858–866.
[45] S. Yan, H. Wang, Semi-supervised learning by sparse representation, in: Proceedings of SIAM international Conference on Data Mining, 2009, pp. 792–801.
[46] M. Zheng, J.J. Bu, C. Chen, C. Wang, L.J. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process. 20 (5) (2011) 1327–1336.
[47] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341.
[48] L.M. Zhang, S. Chen, L. Qiao, Graph optimization for dimensionality reduction with sparsity constraints, Pattern Recognit. 45 (3) (2012) 1205–1210.
[49] J. Yang, D. Chu, L. Zhang, Y. Xu, J. Yang, Sparse representation classifier steered discriminative projection with applications to face recognition, IEEE Trans. Neural Netw. Learn. Syst. 24 (7) (2013) 1023–1035.
[50] Y. Chen, Z. Jin, Reconstructive discriminant analysis: a feature extraction method induced from linear regression classification, Neurocomputing 87 (15) (2012) 41–50.
[51] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, S. Ji, Discriminant sparse neighborhood preserving embedding for face recognition, Pattern Recognit. 45 (8) (2012) 2884–2893.
[52] H. Li, J. Tao, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Netw. 17 (1) (2006) 157–165.
[53] F. Zang, J. Zhang, Discriminative learning by sparse representation for classification, Neurocomputing 74 (2011) 2176–2183.
[54] L. Wei, F. Xu, A. Wu, Weighted discriminative sparsity preserving embedding for face recognition, Knowl.–Based Syst. 57 (2) (2014) 136–145.
[55] E.J. Candes, M.B. Wakin, An introduction to compressive sampling, IEEE Signal Process. Mag. 25 (2) (2008) 21–30.
[56] D.L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (4) (2006) 1289–1306.
[57] F.S. Samaria, A.C. Harter, Parameterisation of a stochastic model for human face identification, in: Proceedings of the Second IEEE Workshop on Applications of Computer Vision, Sarasota, FL, 1994, pp. 138–142.
[58] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 643–660.
[59] A. Martinez, R. Benavente, The AR Face Database, Purdue University, Computer Vision Center, West Lafayette, 1998, CVC technical report 24 [R].

Songjiang Lou received the B.S. degree from Ningbo University in 2005, and M.S and Ph.D. degrees in Computer Science and Measurement & Technique and Automation Equipment from Harbin Engineering University in 2008 and 2011, respectively. He is currently a lecturer in Taizhou University. His research interests include pattern recognition and machine learning.

Xiaoming Zhao received the B.S degree at mathematics from Zhejiang Normal University in 1990 and the M.S degree at software engineering from Beihang University in 2006. He is currently a professor of department of computer science, Taizhou University, China. His research interests include computer vision, pattern recognition, and machine learning.

Yuelong Chuang received his Ph.D. degree in College of Computer Science at Zhejiang University, China. He received his M.S. degree in School of Computer and Communication Engineering in Liaoning Shihua University, China. His research interests include computer vision, machine learning, and pattern recognition.

Haitao Yu received master degree of Computer Science and Technology from Harbin Engineering University in 2008, and then went to work as a teacher of college of Computer Science and Information Technology in Daqing Normal University. Now he is a Ph.D. candidate of Computer Science and Technology in Harbin Engineering University. His main research interests include underwater acoustic sensor network, computer network security and so on.

Shiqing Zhang received the B.S. degree at electronics and information engineering from Hunan University of Commerce in 2003, the M.S degree at electronics and communication engineering from Hangzhou Dianzi University in 2008, and the Ph.D. degree at school of Communication and Information Engineering, University of Electronic Science and Technology of China, in 2012. Currently, he works as an assistant professor of department of physics and electronics engineering, Taizhou University, China. His main research interests include signal processing, computer vision and pattern recognition.