Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Discriminant sparse neighborhood preserving embedding for face recognition

Jie Gui [a,b], Zhenan Sun [a,*], Wei Jia [b], Rongxiang Hu [b], Yingke Lei [c], Shuiwang Ji [d]

[a] *Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[b] *Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China*
[c] *Electronic Engineering Institute, Hefei Anhui 230037,China*
[d] *Department of Computer Science, Old Dominion University, Norfolk, VA, USA, 23529-0162*

ARTICLE INFO

Article history:
Received 5 January 2011
Received in revised form
5 January 2012
Accepted 8 February 2012
Available online 18 February 2012

Keywords:
Sparse neighborhood preserving embedding
Sparse subspace learning
Discriminant learning
Maximum margin criterion
Discriminant sparse neighborhood preserving embedding
Face recognition

ABSTRACT

Sparse subspace learning has drawn more and more attentions recently. However, most of the sparse subspace learning methods are unsupervised and unsuitable for classification tasks. In this paper, a new sparse subspace learning algorithm called discriminant sparse neighborhood preserving embedding (DSNPE) is proposed by adding the discriminant information into sparse neighborhood preserving embedding (SNPE). DSNPE not only preserves the sparse reconstructive relationship of SNPE, but also sufficiently utilizes the global discriminant structures from the following two aspects: (1) maximum margin criterion (MMC) is added into the objective function of DSNPE; (2) only the training samples with the same label as the current sample are used to compute the sparse reconstructive relationship. Extensive experiments on three face image datasets (Yale, Extended Yale B and AR) demonstrate the effectiveness of the proposed DSNPE method.

1. Introduction

In the past two decades, appearance-based face recognition has attracted considerable interests in computer vision and pattern recognition [1,2]. It is well known that the dimension of face images is usually very high. For example, a 100-by-100 pixel face image can be viewed as a 10,000-dimensional vector. High dimensionality of feature vector has become a critical problem in practical pattern recognition applications. The data in the high-dimensional space is usually redundant and may degrade the performance of pattern classifiers when the number of training samples is much smaller than the dimensionality of the input data. A common way to solve these problems is to adopt dimensionality reduction methods. So far, an enormous volume of literature has been devoted to investigate various data-dependent dimensionality reduction methods for projecting the high-dimensional data into low-dimensional feature spaces. These traditional dimensionality reduction methods can be classified into four categories as follows.

The first category is linear dimensionality reduction algorithm (also named as subspace learning algorithm), among which principal component analysis (PCA) and linear discriminant analysis (LDA) are two of the most popular ones [2,3]. Generally, PCA projects the original data into a low-dimensional space which is spanned by the eigenvectors associated with the largest eigenvalues of the covariance matrix of all the data points. However, PCA does not take into consideration the label information of the input data. As a result, PCA will probably lose much useful information which is critical for pattern classification tasks [4]. Unlike PCA, LDA is a supervised method which takes full consideration of the class labels for patterns. It is generally believed that the class information can make the recognition algorithm more discriminative. Thus, LDA has been shown to be more effective than PCA in many applications. One limitation of PCA and LDA is that they only exploit the linear global Euclidean structure. Recent research shows that the face images may reside on a nonlinear submanifold [5,6], which makes PCA and LDA inefficient. In order to overcome the problem, many nonlinear feature extraction methods such as kernel-based approaches and manifold learning-based ones have been developed.

The second category is the kernel-based algorithm [7,8], which uses a linear classifier algorithm to solve non-linear problems by mapping the original non-linear observations into a higher-dimensional space. It is based on the assumption that the nonlinear structure data will be linearly separable in the kernel space. The most popular kernel methods are kernel principal component analysis (KPCA) [8] and kernel Fisher discriminant analysis (KFDA)

[*] Corresponding author. Tel.: +86 10-82610278
E-mail address: znsun@nlpr.ia.ac.cn (Z. Sun)

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.02.005

[7], which are the kernel versions of PCA and LDA. KPCA and KFDA have been proved to be effective in some real world applications. However, the choice of the kernel, which is crucial to the success of these algorithms, has been traditionally entirely left to the user. So many research works are conducted on multiple kernel learning to solve the problem of kernel determination [9].

The third category is manifold learning-based algorithm, which is based on the idea that the data points are actually samples from a low-dimensional manifold that is embedded in a high-dimensional space. The representative algorithms include locally linear embedding (LLE) [5], isometric feature mapping (ISOMAP) [10], Laplacian eigenmaps (LE) [11], Hessian-based locally linear embedding (HLLE) [12], maximum variance unfolding (MVU) [13,14], local tangent space alignment (LTSA) [15,16], Riemannian manifold learning (RML) [17,18], and local spline embedding (LSE) [19], etc. Each manifold learning algorithm attempts to preserve a different geometrical property of the underlying manifold. Local approaches such as LLE and LE, the first step of which is graph construction based on k-nearest-neighbor and ε-ball based methods, aim to preserve the locality proximity relationship among the data, while global approaches like ISOMAP aim to preserve the metrics at all scales. These nonlinear methods do yield impressive results on some benchmark artificial and real world data sets due to their nonlinear nature, geometric intuition, and computational feasibility. However, all these manifold learning algorithms have the out of sample problem [20]. The reason is that they can only yield an embedding of the training data set. Nevertheless, when applied to a new sample, they cannot easily find the sample's image in the embedding space by utilizing the low-dimensional embedding results of the training data set because of the implicitness of the nonlinear map. Thus a dozen of methods have been proposed to solve this problem, e.g., incremental manifold learning [21], low-rank matrix approximation [22], locality preserving projections (LPP) [23], discriminant locality preserving projections based on maximum margin criterion (DLPP/MMC) [24], null space discriminant locality preserving projections (NDLPP) [25], locality preserving discriminant projections (LPDP), etc. [26].

The last one is matrix and tensor embedding algorithm [27,28] which represents patterns as matrixes or high-order tensors instead of vectors. The aforementioned subspace learning algorithms, kernel based algorithm and manifold learning-based algorithm all consider a vector representation of samples. However, the extracted features from many real world vision problems may contain higher-order structure. For example, a captured image is a second-order tensor, i.e., a matrix, and sequential data such as video sequences for event analysis is in the form of a third-order tensor. Thus it is necessary to derive the multilinear forms of these traditional linear feature extraction methods to handle the data as tensors directly. Recently this research field has received a lot of attention from the image processing and computer vision community, and these methods [28–34] have been shown to be much more efficient than the traditional vector-based methods.

Recently, some new methods integrating the theory of sparse representation, compressed sensing and subspace learning (linear dimensionality reduction methods) have been proposed, and have been successfully applied in many practical applications [35–37, 61, 62]. Sparse subspace learning (SSL) [38] is a special family of dimensionality reduction methods which consider "sparsity". It has either of the following two characteristics: (1) finding a subspace spanned by sparse base vectors. The sparsity is enforced on the projection vectors and associated with the feature dimension. The representative methods include sparse principal component analysis (SPCA) [39], sparse nonnegative matrix factorization [36], and nonnegative sparse PCA [40], etc. (2) Aiming at the sparse reconstructive weight which is associated with the sample size. The representative methods include sparse neighborhood preserving embedding (SNPE) [41]. In fact, SNPE is identical to sparsity preserving projections (SPP) [42], which has achieved higher recognition rates than PCA and neighborhood preserving embedding (NPE) for face recognition. Zhang et al. [43] proposed a sparse representation-based classifier (SRC) [35] oriented unsupervised dimensionality reduction algorithm which combines SRC and PCA in its objective function. Yang and Chu [44] proposed the SRC steered discriminative projection (SRC-DP). The basic idea of SRC-DP is to seek a linear transformation such that in the transformed low-dimensional space, the within-class reconstruction residual is as small as possible and simultaneously the between-class reconstruction residual is as large as possible.

However, SNPE suffers from a limitation that it does not encode discriminant information, which is very important for recognition tasks. In this paper, we propose a discriminant sparse neighborhood preserving embedding (DSNPE) algorithm by combining SNPE and maximum margin criterion (MMC) methods, which can be viewed as a new algorithm integrating Fisher criterion and sparsity criterion. It is well known that MMC is a method proposed to maximize the trace of the difference of the between-class scatter matrix and within-class scatter matrix from which LDA can be derived by incorporating some constraints. Thus, DSNPE is proposed by introducing MMC into the objective function of SNPE, which has two advantages: (1) it retains the sparsity characteristic of SNPE; (2) it emphasizes the discriminative information by incorporating MMC, which can make the class mean vectors have a wide spread and make every class scatter in a small space. Furthermore, to further increase the discriminative power of DSNPE, we integrate additional discriminant information. More concretely, to compute the sparse reconstructive relationship, we only use the training samples with the same label as the current sample instead of using all of the training samples. The reason behind this decision is based on the following observation: taking face images into account, the most compact expression of a certain face image is generally given by the face images from the same class [35]. The proposed method is applied to face biometrics and is examined using the Yale, Extended Yale B, and AR face image databases. Experimental results show that it is more suitable for recognition tasks than SNPE.

The remainder of this paper is organized as follows: In Section 2 we will introduce our DSNPE method in details. A theoretical analysis of DSNPE is given in Section 3. The experimental results for applying our method to face recognition will be presented in Section 4, followed by the conclusions in Section 5.

2. Discriminant sparse neighborhood preserving embedding (DSNPE)

2.1. Graph construction based on sparse representation

Instead of considering k-nearest-neighbor and ε-ball based methods as in typical graph construction, we attempt to automatically construct a graph G and make it well preserve the discriminative information based on sparse representation (SR). In the past few years, SR has received a great deal of attentions, which was initially proposed as an extension of traditional signal processing methods such as Fourier and wavelet. The problem solved by sparse representation is to search for the most compact representation of a signal in terms of linear combination of patterns in an over-complete dictionary. SR has been successfully used in image super-resolution [45,46], image denoising [47–49], signal reconstruction [50], signal recovery [51], etc.

SR has compact mathematical expression. Given a signal (or an image with vector pattern) $x \in R^D$, and a matrix $X=[x_1,x_2,\ldots,x_n] \in R^{n\times D}$

containing the elements of an over-complete dictionary [52] in its columns, the goal of SR is to represent x using as few entries of X as possible. The objective function can be described as follows:

$$\min_{s_i} \quad \|s_i\|_0$$
$$\text{s.t.} \quad x_i = Xs_i \tag{1}$$

or

$$\min_{s_i} \quad \|s_i\|_0$$
$$\text{s.t.} \quad \|x_i - Xs_i\| \prec \varepsilon \tag{2}$$

where $s_i = [s_{i,1},\ldots,s_{i,i-1},0,s_{i,i+1},\ldots,s_{in}]^T$ is an n-dimensional vector in which the ith element is equal to zero (implying that the x_i is removed from X), and the elements $s_{i,j}, j \neq i$ denote the contribution of each x_j to reconstructing x_i. Unfortunately, this criterion is not convex, and finding the sparsest solution of Eq. (1) is NP-hard. This difficulty can be bypassed by convexizing the problem and using l_1 instead of l_0. The l_1 minimization problem can be solved by LASSO [53] or LARS [54]. After repeating l_1 minimization problem to all the points, the sparse weight matrix can be expressed as $S = [s_1,\ldots,s_n]^T$. Then, the new constructed graph is $G = \{X,S\}$, where X is the training sample set and S is the edge weight matrix.

In the following, we give two reasons why SR is more suitable to graph construction than k-nearest-neighbor and ε-ball based methods.

(1) *Parameter-free*. SR does not need to determine the model parameters such as the neighborhood size k of k-nearest-neighbor and ε of ε-ball based methods, which are generally difficult to set in practice. In contrast, the advantage of being parameter-free makes SR easy to use in practice. In fact, the data distribution probability may vary greatly at different areas of the data space, which results in distinctive neighborhood structure for each instance. However, both k-nearest-neighbor and ε-ball based methods use a predefined parameter to determine the neighborhoods for all the data. It seems to be unreasonable that all data points share the same parameter for k-nearest-neighbor and ε-ball based methods, which may not characterize the manifold structure well, especially in under sampling case. Obviously, compared to k-nearest-neighbor and ε-ball based methods, SR has the merit of being parameter-free.

(2) *Robustness to data noise*. The data noise is inevitable especially for visual data, and the robustness is a desirable property for a satisfying graph construction method. The graph constructed by k-nearest-neighbor and ε-ball based methods is based on pair-wise Euclidean distance, which is very sensitive to data noise. It means that the graph structure is easy to change when unfavorable noise comes in. However, SR has been shown to be robust to data noise in [35].

Table 1
Discriminant sparse neighborhood preserving embedding.

Input: training set $X = \{(x_i, y_i)\}_{i=1}^{N}$
Output: $D \times d$ feature matrix w extracted from X
1. Project the image set $\{x_i\}$ into the PCA subspace by throwing away the smallest principal components
2. Construct weight matrix S using Eqs. (15) or (16)
3. Perform eigenvalue decomposition using Eq. (20), construct $D \times d$ feature matrix w whose columns consist of the eigenvectors corresponding to its d smallest eigenvalues.

2.2. Sparse neighborhood preserving embedding (SNPE)

In [41,42], the objective function of SNPE is defined as

$$\min \quad \sum_{i=1}^{n} \|w^T x_i - w^T X s_i\|^2 \tag{3}$$

where w is the projection matrix. By some simple algebra formulations (see the appendix), the objective function can be reduced to

$$\sum_{i=1}^{n} \|w^T x_i - w^T X s_i\|^2 = w^T X(I - S - S^T + S^T S)X^T w \tag{4}$$

For compact expression, the objective function can be further transformed to an equivalent form as follows:

$$\sum_{i=1}^{n} \|w^T x_i - w^T X s_i\|^2 = w^T X(I - S - S^T + S^T S)X^T w = w^T X S_\alpha X^T w \tag{5}$$

where $S_\alpha = I - S - S^T + S^T S$

In addition, to avoid degenerate solutions, a constraint is added

$$w^T X X^T w = I \tag{6}$$

Therefore, the minimization problem is reduced to

$$\arg \quad \min w^T X S_\alpha X^T w$$
$$\text{s.t.} \quad w^T X X^T w = I \tag{7}$$

Therefore, the transformation matrix that minimizes the objective functions is given by the minimum eigenvalues solution to the generalized eigenvalues problem

$$X S_\alpha X^T w = \lambda X X^T w \tag{8}$$

It is easy to show that the matrices $X S_\alpha X^T$ and XX^T are symmetric and positive semidefinite. The vectors w_i that minimize the objective function are given by minimum eigenvalues solutions to the generalized eigenvalues problem. Let the column vectors $w_0, w_1, \ldots, w_{d-1}$ be the solutions of Eq. (8), ordered according to their eigenvalues, $\lambda_0, \lambda_1, \ldots, \lambda_{d-1}$.Thus, the embedding is written as follows:

$$x_i \to y_i = w^T x_i, \quad w = [w_0, w_1, \ldots, w_{d-1}] \tag{9}$$

where y_i is a d-dimensional vector, and w is a $D \times d$ matrix.

2.3. Maximizing margin criterion (MMC)

Maximum margin criterion (MMC) [55,56] is proposed to maximize the (average) margin between classes after dimensionality reduction. MMC can represent class separability better than PCA. Furthermore, LDA can be derived from MMC by incorporating some constraints. However, MMC does not suffer from the small sample size problem, which is known to cause serious stability problems for LDA.

In [55,56], the objective function of MMC is written as

$$J_1 = \max\left\{ \sum_{ij} p_i p_j (d(m_i, m_j) - s(m_i) - s(m_j)) \right\} \tag{10}$$

where p_i and p_j are the prior probability of class i and class j, m_i and m_j are the mean vectors of class i and class j. Here $d(m_i, m_j)$, $s(m_i)$, and $s(m_j)$ are defined as

$$d(m_i, m_j) = \|m_i - m_j\| \tag{11}$$

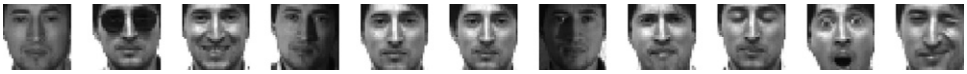

Fig. 1. Eleven cropped and resized samples of one person in Yale face database.

$$s(m_i) = \mathrm{tr}(S_i) \tag{12}$$

$$s(m_j) = \mathrm{tr}(S_j) \tag{13}$$

where S_j is the covariance matrix of class j.

Thus the optimized function can be derived as follows:

$$J_2 = \max \mathrm{tr}(S_b - S_w) \tag{14}$$

The matrix S_b is called the between-class scatter matrix and S_w is called the within-class scatter matrix.

2.4. Discriminant sparse neighborhood preserving embedding (DSNPE)

In this section, we will discuss the solution of DSNPE. Inspired by the observation that the most compact expression of a certain face image is generally given by the face images from the same class [35], we modify the original sparse representation as

$$\begin{aligned} &\min_{s_i} \quad \|s_i\|_1 \\ &\text{s.t.} \quad x_i = X_k s_i \quad \mathrm{label}(x_i) = k \end{aligned} \tag{15}$$

or

$$\begin{aligned} &\min_{s_i} \quad \|s_i\|_1 \\ &\text{s.t.} \quad \|x_i - X_k s_i\| \prec \varepsilon \quad \mathrm{label}(x_i) = k \end{aligned} \tag{16}$$

where X_k denote the set of training samples whose label is the same as x_i. That is to say, to compute the sparse reconstructive relationship, we only use the training samples with the same label as the current sample instead of using all of the training samples.

Furthermore, if a linear transformation $Y = w^T X$ can maximize J_2, an optimal subspace for pattern classification will be explored. This is because the linear transformation aims to project a pattern closer to patterns in the same class but farther from those in different classes, which is exactly the goal for classification. That is to say, to find an optimal linear subspace for classification means to maximize the following optimized function:

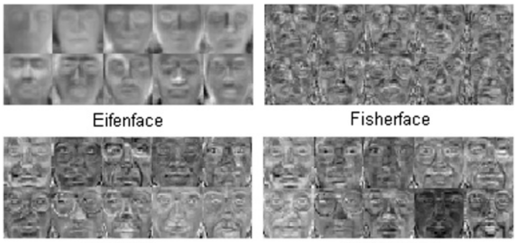

Fig. 2. Top 10 Eigenfaces, Fisherfaces, DSNPE1faces and DSNPE2faces of Yale dataset.

$$J_3 = \max \mathrm{tr}(w^T (S_b - S_w) w) \tag{17}$$

If the linear transformation obtained by SNPE can satisfy J_3 simultaneously, the discriminability of the data will be improved greatly. Thus the solution for DSNPE can be represented as the following multi-object optimization problem:

$$\begin{aligned} &\begin{cases} \min \mathrm{tr}(w^T X S_\alpha X^T w) \\ \max \mathrm{tr}(w^T (S_b - S_w) w) \end{cases} \\ &\text{s.t.} \quad w^T X X^T w = I \end{aligned} \tag{18}$$

The solution to the constrained multi-object optimization problem is to find a subspace which preserves the sparsity property and maximizes the margin between different classes simultaneously, so it can be changed into the following constrained problem:

$$\begin{aligned} &\min \ \mathrm{tr}(w^T (X S_\alpha X^T - \gamma (S_b - S_w)) w) \\ &\text{s.t.} \quad w^T X X^T w = I \end{aligned} \tag{19}$$

where γ is a parameter to balance the sparsity and the discriminant information.

Eq. (19) can be solved by Lagrangian multiplier method:

$$\frac{\partial}{\partial w} \ \mathrm{tr}(w^T (X S_\alpha X^T - \gamma (S_b - S_w)) w - \lambda_i (w^T X X^T w - I)) = 0$$

where λ_i is the Lagrangian multiplier. Then, we can get

$$(X S_\alpha X^T - \gamma (S_b - S_w)) w_i = \lambda_i X X^T w_i \tag{20}$$

where w_i is the generalized eigenvector of $X S_\alpha X^T - \gamma (S_b - S_w)$ and XX^T; λ_i is the corresponding eigenvalue.

Let the column vectors $w_0, w_1, \ldots, w_{d-1}$ be the solutions of Eq. (20), ordered according to their first d smallest eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{d-1}$. Thus, the embedding is written as follows:

$$x_i \rightarrow y_i = w^T x_i, \quad w = [w_0, w_1, \ldots, w_{d-1}]$$

where y_i is a d-dimensional vector and w is a $D \times d$ matrix.

The main procedure for the discriminant sparse neighborhood preserving embedding algorithm is summarized in Table 1.

2.5. Time complexity analysis

In this section, we theoretically analyze the time complexity of our algorithm. We omit the time complexity analysis of sparse learning, because there are a number of software packages to realize the algorithm of sparse learning and different package has different time complexities. For convenience, we give a notation that the number of principal components in the PCA step of DSNPE is q. The DSNPE contains the PCA step and the eigendecomposition step using Eq. (20). Since the PCA step of DSNPE is

Table 2
The maximal average recognition rates (percent) across 20 runs on the Yale database and the corresponding standard deviations (std) and dimensions (shown in parentheses).

Method	2 Train	3 Train	4 Train	5 Train	6 Train	7 Train	8 Train
Baseline	42.63 ± 3.79(1024)	48.08 ± 4.28(1024)	52.86 ± 4.19(1024)	55.44 ± 3.86(1024)	58.80 ± 5.28(1024)	59.67 ± 5.29(1024)	63.44 ± 5.47(1024)
PCA	42.63 ± 3.79(29)	48.08 ± 4.28(44)	52.86 ± 4.19(59)	55.44 ± 3.86(74)	59.13 ± 5.29(30)	59.83 ± 6.14(33)	64.33 ± 5.70(50)
LPP	57.19 ± 5.51(14)	67.92 ± 4.25(14)	75.14 ± 5.46(16)	77.22 ± 3.50(14)	81.6 ± 4.94(14)	82.25 ± 4.69(14)	84.11 ± 5.21(15)
NDLPP	56.11 ± 5.28(14)	69.70 ± 3.66(14)	77.47 ± 4.60(14)	81.77 ± 3.71(14)	84.60 ± 3.83(14)	87.41 ± 3.91(14)	89.88 ± 3.70(14)
LPDP	56.74 ± 5.90(14)	71.75 ± 4.50(14)	78.90 ± 3.86(16)	81.78 ± 3.75(13)	86.73 ± 3.95(14)	88.17 ± 3.24(14)	90.67 ± 2.35(14)
DLPP/MMC	58.19 ± 5.85(14)	70.08 ± 4.4(15)	78.14 ± 4.28(14)	83.56 ± 4.17(18)	85.53 ± 3.76(16)	88.33 ± 3.94(14)	89.56 ± 4.17(18)
LDA	45.19 ± 5.10(10)	59.42 ± 4.62(13)	68.95 ± 5.87(13)	74.89 ± 3.52(14)	79.27 ± 4.69(14)	79.83 ± 6.73(13)	83.22 ± 5.51(14)
SNPE1	66.77 ± 4.16(27)	69.95 ± 2.30(41)	73.61 ± 2.99(55)	74.27 ± 3.71(74)	77.86 ± 4.82(85)	76.91 ± 6.19(101)	79.33 ± 6.95(114)
SNPE2	66.14 ± 4.48(28)	70.29 ± 3.64(43)	73.57 ± 3.71(56)	73.77 ± 5.26(67)	78 ± 5.0910(86)	77.41 ± 5.81(96)	81.44 ± 6.04(110)
DSNPE1	72.33 ± 5.86(29)	**82.33 ± 3.58(44)**	**86.85 ± 2.90(59)**	**90.61 ± 2.98(74)**	**93.60 ± 2.65(89)**	**93.41 ± 3.17(104)**	**96.00 ± 3.10(119)**
DSNPE2	**72.40 ± 6.38(29)**	80.70 ± 3.76(44)	86.66 ± 3.12(59)	89.88 ± 2.76(74)	92 ± 2.59(89)	92.58 ± 4.02(104)	95 ± 3.13(119)

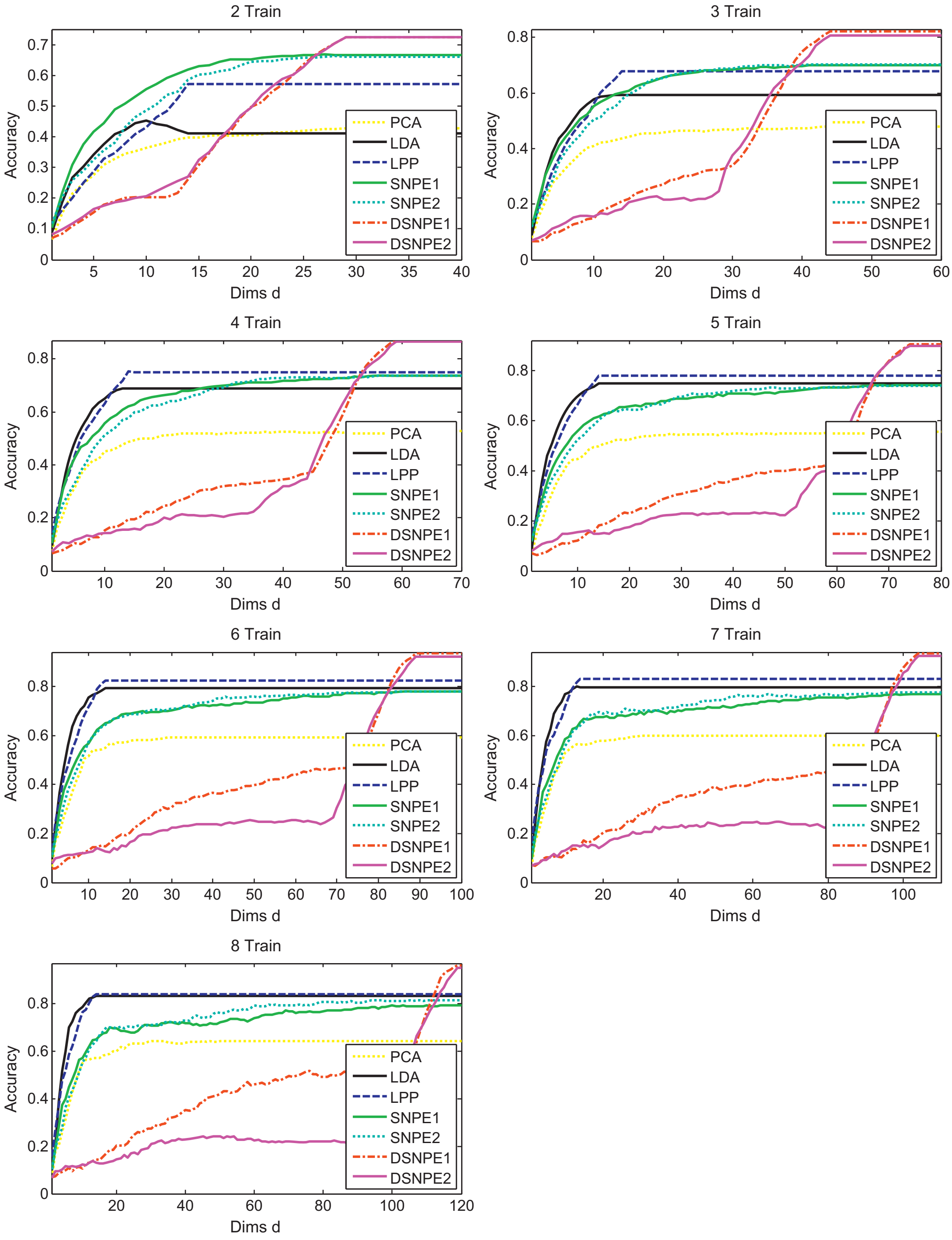

Fig. 3. Recognition accuracy vs. dimensionality on Yale database with 2,3,4,5,6,7,8 images for each individual randomly selected for training.

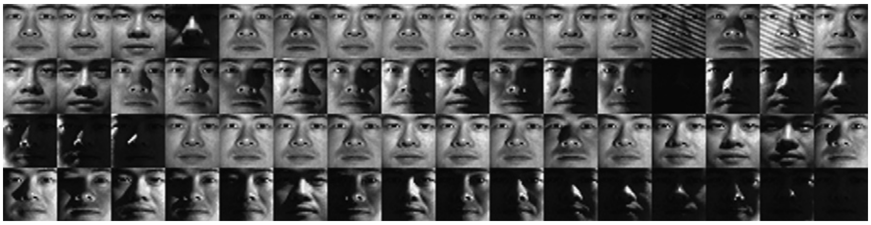

Fig. 4. Thirty two cropped and resized samples of one person in Extended Yale B face database.

the same as the one often used in the other algorithms such as the classical LDA (i.e., PCA+LDA) and LPP, we focus on the time complexity of the eigendecomposition step using Eq. (20), which is $o(q^3)$. Hence, the time complexity of DSNPE is $o(q^3)$.

3. Theoretical analysis of DSNPE

In this section, we give some theoretical analyses to better reveal the characteristic of DSNPE. At first, a lemma is presented as follows:

Lemma 1. [57]. *For symmetric matrix* $A\in\mathbb{R}^{n\times n}$, $E\in\mathbb{R}^{n\times n}$, *let* $A=Q\Lambda Q^T$ *be the eigen-decomposition of A and* $A+E=B=P\Lambda_1 P^T$ *be the eigen-decomposition of B. Write* $Q=[q_1,q_2,\ldots,q_n]$, $P=[p_1,p_2,\ldots,p_n]$, *where* q_i *and* p_i *are the normalized eigenvectors of A and B, respectively. Let* θ *denote the acute angle between* q_i *and* p_i, *then*

$$\sin(\theta)\le\alpha\|E\|_2$$

where α is a constant that only depends on A.

Table 3
The maximal average recognition rates (percent) across 20 runs on the Extended Yale B database and the corresponding standard deviations (std) and dimensions (shown in parentheses).

Methods	5 Train	10 Train
Baseline	36.55 ± 1.54	53.43 ± 0.82
PCA	36.55 ± 1.55(188)	53.43 ± 0.82(372)
LPP	59.14 ± 2.58(189)	69.40 ± 2.16(379)
LDA	75.14 ± 1.78(37)	87.22 ± 1.09(37)
NDLPP	77.28 ± 1.84(37)	87.66 ± 1.08(37)
LPDP	73.77 ± 1.90(155)	87.96 ± 0.73(197)
DLPP/MMC	73.66 ± 1.86(41)	87.65 ± 1.05(74)
SNPE1	77.24 ± 1.84(189)	82.69 ± 2.28(379)
SNPE2	75.58 ± 1.6701(189)	84.54 ± 1.26(379)
DSNPE1	**79.20 ± 1.85(189)**	**88.17 ± 0.85(379)**
DSNPE2	77.67 ± 1.86(189)	86.95 ± 1.35(379)

The following Theorem 1 characterizes the solution of DSNPE when the parameter γ is approaching infinity. Note that Theorem 1 requires the positive definiteness of XX^T, which always holds for our algorithm since we use PCA to preprocess the data. Without loss of generality, we further assume the data matrix X has been centered.

Theorem 1. *When* $\gamma\to\infty$, *the* w_i *obtained by the proposed DSNPE method converges to the generalized eigenvector* m_i *of the between-class scatter matrix* S_b *and the within-class scatter matrix* S_w, *i.e., when* $\gamma\to\infty$, *there exists a constant* β *such that*

$$w_i-\beta m_i=0$$

Proof. If the both sides of Eq. (20) are divided by γ, we have

$$\left(\frac{1}{\gamma}XS_\alpha X^T-(S_b-S_w)\right)w_i=\frac{\lambda_i}{\gamma}XX^Tw_i \tag{21}$$

Eq. (21) is equivalent to

$$(XX^T)^{-0.5}\left(\frac{1}{\gamma}XS_\alpha X^T-(S_b-S_w)\right)(XX^T)^{-0.5}(XX^T)^{0.5}w_i=\frac{\lambda_i}{\gamma}(XX^T)^{0.5}w_i \tag{22}$$

Therefore $(XX^T)^{0.5}w_i$ is the eigenvector of $(XX^T)^{-0.5}((1/\gamma)XS_\alpha X^T-(S_b-S_w))(XX^T)^{-0.5}$.

On the other hand, from the definition of m_i, we obtain

$$S_bm_i=c_iS_wm_i \tag{23}$$

where c_i is the corresponding eigenvalue of m_i.

Since X has been centered, then we have $XX^T=nS_t=n(S_b+S_w)$. With (23), we then have

$$-(XX^T)^{-0.5}(S_b-S_w)(XX^T)^{-0.5}(XX^T)^{0.5}m_i=d_i(XX^T)^{0.5}m_i \tag{24}$$

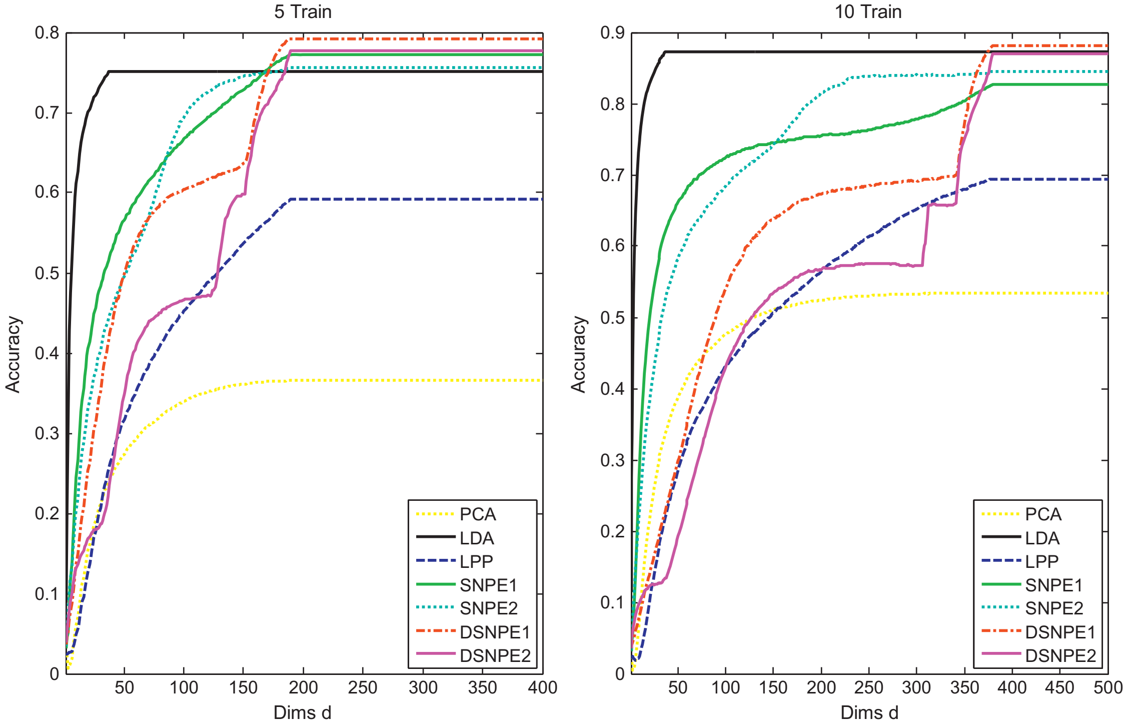

Fig. 5. Recognition accuracy vs. dimensionality on Extended Yale B database with 5, 10 images for each individual randomly selected for training.

where $d_i = -1 + c_i/n + nc_i$. Therefore $(XX^T)^{0.5} m_i$ is the eigenvector of $-(XX^T)^{-0.5}(S_b - S_w)(XX^T)^{-0.5}$.

Let θ denote the acute angle between $(XX^T)^{0.5}$ m_i and $(XX^T)^{0.5}$ w_i. By directly applying Lemma 1 we have

$$\sin(\theta) \le \frac{\alpha}{\gamma} \|(XX^T)^{-0.5}(XS_\alpha X^T)(XX^T)^{-0.5}\|_2.$$

when γ goes to infinity, it is easy to see that $\sin(\theta) \to 0$. Therefore, there exists a constant β such that

$$(XX^T)^{0.5}(w_i - \beta m_i) = 0 \qquad (25)$$

Since we assume that $XX^T \succ 0$, we have

$$w_i - \beta m_i = 0 \qquad (26)$$

This completes the proof of Theorem 1. □

When the parameter γ assumes the value of zero, DSNPE degenerates into SNPE. From this point it can be concluded that SNPE is a special case of DSNPE.

4. Experimental results

In this section, we conducted a set of experiments to verify the effectiveness of the proposed DSNPE method. Three face databases were used, including Yale, Extended Yale B and the AR face image database.

In each experiment, the image set was partitioned into a training set and test set with different numbers. For ease of representation, the experiments were named as p-train, which means that p images per individual were selected for training and the remaining images for test. To robustly evaluate the performance of different algorithms in different training and testing conditions, we selected images randomly and repeated the experiment 20 times in each condition. We exhibited the results in the form of mean recognition rate with standard deviation.

We compare DSNPE with several representative dimensional reduction methods such as PCA [2], LDA [2], LPP [23,58], DLPP/MMC [24], LPDP [26], NDLPP [25], SNPE1 [42], and SNPE2 [42]. The nearest neighbor classifier is employed for classification. For LPP, the number of nearest neighbors k is taken to be $p-1$ as done in [59] where p is the number of images per individual selected for training. For DSNPE, we simply set the value of γ as 1.

4.1. Experimental results on the Yale database

The Yale face database was constructed at the Yale Center for Computation Vision and Control. There are 165 images of 15 individuals (each person providing 11 different images). The images demonstrate variations in lighting condition (left-light, center-light and right-light), facial expression (normal, happy, sad, sleepy, surprised, and wink), and with or without glasses. All images were also in grayscale and cropped and resized to the resolution of 32 × 32 pixels. We pre-processed the data by normalizing each face vector to the unit. Shown in Fig. 1 is one object from Yale database. The top ten Eigenfaces, Fisherfaces, DSNPE1faces, and DSNPE2faces of Yale images are shown in Fig. 2.

For each person, p images (p varying from 2 to 8) were randomly selected for training, and the rest were used as test samples. The training set was used to learn a face subspace. Recognition was then performed in the subspaces. In general, the recognition rates vary with the dimension of the face subspace. Table 2 shows the maximal average recognition rates across 20 runs of each method under nearest neighbor classifier and their corresponding standard deviations (std) and dimensions, where the best results are highlighted in bold.

The recognition rate curves of different algorithms are drawn in Fig. 3 where SNPE1 denotes the SNPE algorithm based on Eq. (1), SNPE2 denotes the SNPE based on Eq. (2).

DSNPE1 denotes the DSNPE algorithm based on Eq. (15) and DSNPE2 denotes the DSNPE based on Eq. (16). Due to the space limitation, we only draw the recognition rate curves of some representative methods in Fig. 3. For the baseline method, we simply performed face recognition in the original 1024-dimensional image space. Note that the upper bound of the dimensionality of LDA is $c-1$ where c is the number of individuals [2].

As can be seen, our algorithm DSNPE1 outperformed all other methods except for 2 train while the PCA method performed the worst in all cases. It is very interesting that the PCA method and the baseline method have the same performance when p varies from 2 to 5 which is consistent with the results in many publications such as [60].

4.2. Experimental results on the Extended Yale B database

The Extended Yale B database [50] contains 2414 front-view face images of 38 individuals. For each individual, about 64 pictures were taken under various laboratory-controlled lighting conditions. In our experiments, we use the cropped images with the resolution of 32 × 32. We pre-processed the data by normalizing each face vector to the unit. Thirty two cropped sample images of one person in the Extended Yale B database after the scale normalization are displayed in Fig. 4.

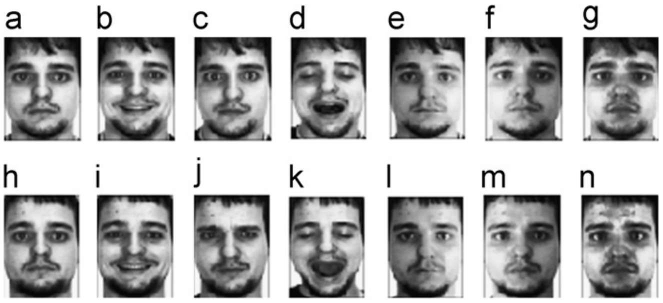

Fig. 6. Some typical samples of the cropped images found in the AR face image database.

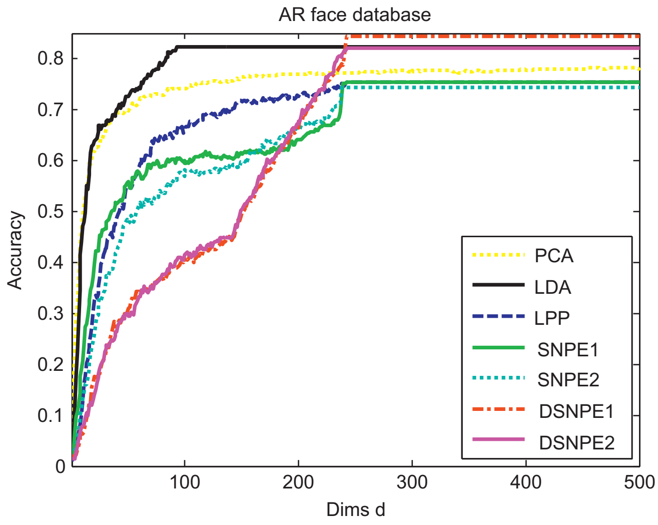

Fig. 7. Recognition accuracy vs. dimensionality on the AR face database.

Table 4
Maximal recognition rates (percent) on the AR face database and the corresponding dimensions.

Methods	Baseline	PCA	LPP	LDA	NDLPP	LPDP	DLPP/MMC	SNPE1	SNPE2	DSNPE1	DSNPE2
Recognition rate	78.14	78.14	75.42	82.42	83.43	83.43	84	75.42	74.28	**84.28**	82
Dimension	2520	455	243	93	99	229	70	243	238	243	243

For each individual, p images (p equals to 5 or 10) were randomly selected for training and the rest were used for test. The experimental design is the same as in Section 4.1. The maximal average recognition rate, the corresponding dimensionality and the standard deviations across 20 runs of tests of each method are shown in Table 3. The best results are highlighted in bold face font. In addition, we draw the recognition rate curves of some representative algorithms in Fig. 5.

As can be seen, our DSNPE algorithm performed the best in all cases. Moreover, the PCA method and the baseline method have nearly the same performance as the Yale face database.

4.3. *Experimental results on the AR face database*

The last experiment was tested using the AR face database (http://rvl1.ecn.purdue.edu/∼aleix/aleix_face_DB.html) which contains over 4000 color face images of 126 individuals (70 men and 56 women), including frontal views of faces with different facial expressions, illumination conditions and occlusions. The pictures of most persons were taken in two sessions (separated by two weeks). Each section contains 13 color images. In our experiments here, we use a subset of the AR face database provided and preprocessed by Martinez [4]. This subset contains 1400 face images corresponding to 100 person (50 men and 50 women), where each person has 14 different images with illumination change and expressions. The original resolution of these image faces is 165 × 120. Here, for computational convenience, we manually cropped the face portion of the image and then normalized it to 62 × 44 pixels. The normalized images of one person are shown in Fig. 6, where the images in the first row in Fig. 6 are from Session 1, and the images in the second row are from Session 2. The details of the images are: Fig. 6a neutral expression, Fig. 6b smile, Fig. 6c anger, Fig. 6d scream, Fig. 6e left light on; Fig. 6f right light on, Fig. 6g all sides light on, and Fig. 6h–n were taken under the same conditions as Fig. 6a–g. Since AR database has naturally been partitioned into two sessions, we also consider this case in our experiments. In this experiment, images from the first session (i.e., Fig. 6a–g) were used for training, and images from the second session (i.e., Fig. 6h–n) were used for test.

As the training set and test set are fixed, we only give the recognition rates of different algorithms. The maximal recognition rate of each method and the corresponding reduced dimension are listed in Table 4. The recognition rate curves of some representative algorithms vs. the variation of dimensions are shown in Fig. 7. From the experimental results, we can see that DSNPE1 achieves the highest recognition rate.

5. Conclusions

In this paper, we developed a new sparse dimensionality reduction method called discriminant sparse neighborhood preserving embedding (DSNPE). Our proposed method combines sparsity criterion and maximum margin criterion (MMC) together to project the input high-dimensional image into a low-dimensional feature vector. Therefore both the robustness advantage of sparse representation and distinctiveness advantage of MMC are integrated to develop a good pattern recognition solution.

The proposed DSNPE method is applied for face recognition. The testing results on three face image databases, i.e., Yale, Extended Yale B and AR face database demonstrate that DSNPE is more effective than some popular dimensionality reduction algorithms. These experiments are mainly designed to prove the effectiveness of DSNPE for pattern recognition task oriented feature dimensionality reduction. Because the DSNPE algorithm is directly applied on the original facial images rather than local features such as Gabor, LBP features, the accuracy of face recognition in our experiments has still a big gap from practical face recognition applications. However, we argue that DSNPE provides a useful tool of dimensionality reduction which may benefit state-of-the-art pattern recognition algorithms. Our future work will apply DSNPE algorithm on advanced visual features rather than the original pixel intensity values. And some useful strategies (e.g., localizing the training images, synthesizing virtual samples) will also be incorporated into the pattern recognition scheme to achieve much higher pattern recognition accuracy for face, iris and palmprint recognition.

Acknowledgments

This work was supported by the National Basic Research Program of China (Grant no. 2012CB316300), the National Science Foundation of China (Grant nos. 61100161, 61175022, 60736018, 60905023, 61075024, and 61005007), the International S&T Cooperation Program of China (Grant no. 2010DFB14110), National Key Technology R&D Program (Grant No. 2012BAK02B01) and the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant nos. Y023A61121 and Y023A11292).

Appendix. The formulation for deriving Eq. (4)

$$
\begin{aligned}
\sum_{i=1}^{n} \|w^T x_i - w^T X s_i\|^2 &= \sum_{i=1}^{n} w^T (x_i - X s_i)(x_i - X s_i)^T w \\
&= w^T \left(\sum_{i=1}^{n} (x_i - X s_i)(x_i - X s_i)^T \right) w \\
&\times w^T \left(\sum_{i=1}^{n} x_i x_i^T - \left(\sum_{i=1}^{n} x_i s_i^T \right) X^T - X \left(\sum_{i=1}^{n} s_i x_i^T \right) \right. \\
&\left. + X \left(\sum_{i=1}^{n} s_i s_i^T \right) X^T \right) w \\
&\times \sum_{i=1}^{n} x_i x_i^T = [x_1 x_2 \cdots x_n] \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = XX^T
\end{aligned}
$$

Similarly, $\sum_{i=1}^{n} s_i s_i^T = S^T S$

$$\left(\sum_{i=1}^{n} x_i s_i^T\right) = [x_1 x_2 \cdots x_n]\begin{bmatrix} s_1^T \\ s_2^T \\ \vdots \\ s_n^T \end{bmatrix} = XS$$

Similarly, $$\left(\sum_{i=1}^{n} s_i x_i^T\right) = [s_1 s_2 \cdots s_n]\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = S^T X^T$$

Thus, $$\sum_{i=1}^{n} \|w^T x_i - w^T X s_i\|^2 = w^T X (I - S - S^T + S^T S) X^T w$$

References

[1] H. Murase, S.K. Nayar, Visual learning and recognition of 3-D objects from appearance, International Journal of Computer Vision 14 (1995) 5–24. (January).

[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 711–720. (July).

[3] J. Gui, S.L. Wang, Y.K. Lei, Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data, Artificial Intelligence in Medicine 50 (2010) 181–191. November).

[4] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 228–233. (February).

[5] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326. (December 22).

[6] H.S. Seung, D.D. Lee, Cognition—the manifold ways of perception, Science 290 (2000) 2268. (December 22).

[7] G. Baudat, F.E. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (2000) 2385–2404. October).

[8] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998) 1299–1319. (July 1).

[9] C. Cortes, M. Mohri, A. Rostamizadeh, Two-stage learning kernel algorithms, in: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), 2010.

[10] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2324. (December 22).

[11] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (2003) 1373–1396. (June).

[12] D.L. Donoho, C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences of the United States of America 100 (2003) 5591–5596. (May 13).

[13] C.P. Hou, Y.Y. Jiao, Y. Wu, D.Y. Yi, Relaxed maximum-variance unfolding, Optical Engineering 47 (2008). (July).

[14] D.K. Saxena, K. Deb, Non-linear dimensionality reduction procedures for certain large-dimensional multi-objective optimization problems: employing correntropy and a novel maximum variance unfolding, Evolutionary Multi-Criterion Optimization, Proceedings 4403 (2007) 772–787.

[15] Z.Y. Zhang, H.Y. Zha, Nonlinear dimension reduction via local tangent space alignment, Intelligent Data Engineering and Automated Learning 2690 (2003) 477–481.

[16] Y.B. Zhan, J.P. Yin, Robust local tangent space alignment, Neural Information Processing, Part 1, Proceedings 5863 (2009) 293–301.

[17] T. Lin, H.B. Zha, Riemannian manifold learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 796–809. May).

[18] T. Lin, H.B. Zha, S.U. Lee, Riemannian manifold learning for nonlinear dimensionality reduction, Computer Vision—ECCV 2006, Part 1, Proceedings 3951 (2006) 44–55.

[19] S.M. Xiang, F.P. Nie, C.S. Zhang, C.X. Zhang, Nonlinear dimensionality reduction with local spline embedding, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1285–1298. September).

[20] Y. Bengio, J.F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, M. Ouimet, Out-of-sample extensions for LLE, isomap, mds, eigenmaps, and spectral clustering, in: Proceedings of Advances in Neural Information Processing Systems 16 (NIPS'03), 2004, pp. 177–184.

[21] M.H.C. Law, A.K. Jain, Incremental nonlinear dimensionality reduction by manifold learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 377–391. March).

[22] D.Y. Yeung, H. Chang, G. Dai, A scalable kernel-based semisupervised metric learning algorithm with out-of-sample generalization ability, Neural Computation 20 (2008) 2839–2861. (November).

[23] X.F. He, S.C. Yan, Y.X. Hu, P. Niyogi, H.J. Zhang, Face recognition using Laplacianfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 328–340. March).

[24] G.F. Lu, Z. Lin, Z. Jin, Face recognition using discriminant locality preserving projections based on maximum margin criterion, Pattern Recognition 43 (2010) 3572–3579. October).

[25] W.G. Gong, L.P. Yang, X.H. Gu, W.H. Li, Y.X. Liang, Null space discriminant locality preserving projections for face recognition, Neurocomputing 71 (2008) 3644–3649. October).

[26] J. Gui, W. Jia, L. Zhu, S.L. Wang, D.S. Huang, Locality preserving discriminant projections for face and palmprint recognition, Neurocomputing 73 (2010) 2696–2707. (August).

[27] S.C. Yan, D. Xu, Q. Yang, L. Zhang, X.O. Tang, H.J. Zhang, Multilinear discriminant analysis for face recognition, IEEE Transactions on Image Processing 16 (2007) 212–220. January).

[28] J. Yang, D. Zhang, A.F. Frangi, J.Y. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 131–137. (January).

[29] X. He, D. Cai, P. Niyogi, Tensor subspace analysis, presented at the Advances in Neural Information Processing Systems, 2006.

[30] Y.T. Wei, H. Li, L.Q. Li, Tensor locality sensitive discriminant analysis and its complexity, International Journal of Wavelets Multiresolution and Information Processing 7 (2009) 865–880. (November).

[31] H.P. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Uncorrelated multilinear discriminant analysis with regularization and aggregation for tensor object recognition, IEEE Transactions on Neural Networks 20 (2009) 103–123. (January).

[32] D.C. Tao, X.L. Li, X.D. Wu, S. Maybank, Tensor rank one discriminant analysis—a convergent method for discriminative multilinear subspace selection, Neurocomputing 71 (2008) 1866–1882. (June).

[33] X.L. Li, S. Lin, S.C. Yan, D. Xu, Discriminant locally linear embedding with high-order tensor data, IEEE Transactions on Systems Man and Cybernetics Part B—Cybernetics 38 (2008) 342–352. (April).

[34] D.C. Tao, X.L. Li, X.D. Wu, S.J. Maybank, General tensor discriminant analysis and Gabor features for gait recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 1700–1715. (October).

[35] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 210–227. (February).

[36] P.O. Hoyer, Non-negative matrix factorization with sparseness constraints, Journal of Machine Learning Research 5 (2004) 1457–1469. (November).

[37] X.R. Pu, Z. Yi, Z.M. Zheng, W. Zhou, M. Ye, Face recognition using Fisher non-negative matrix factorization with sparseness constraints, Advances in Neural Networks—ISNN 2005, Part 2, Proceedings (2005) 112–117.

[38] D. Cai, X. He, J. Han, Spectral regression: a unified approach for sparse subspace learning, presented at the International Conference on Data Mining (ICDM'07), Omaha, NE, 2007.

[39] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, Journal of Computational and Graphical Statistics 15 (2006) 265–286. (June).

[40] R. Zass, A. Shashua, Nonnegative sparse PCA, Advances in Neural Information Processing Systems 19 (2007) 1561.

[41] B. Cheng, J.C. Yang, S.C. Yan, Y. Fu, T.S. Huang, Learning with l(1)-graph for image analysis, IEEE Transactions on Image Processing 19 (2010) 858–866. April).

[42] L.S. Qiao, S.C. Chen, X.Y. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognition 43 (2010) 331–341. January).

[43] L. Zhang, M. Yang, Z. Feng, D. Zhang, On the dimensionality reduction for sparse representation based face recognition, in: 2010 International Conference on Pattern Recognition.

[44] J. Yang, D. Chu, Sparse representation classifier steered discriminative projection, in: 2010 International Conference on Pattern Recognition.

[45] K.I. Kim, Y. Kwon, Single-image super-resolution using sparse regression and natural image prior, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1127–1133. (June).

[46] J.C. Yang, J. Wright, T. Huang, Y. Ma, Image super-resolution as sparse representation of raw image patches, 2008 IEEE Conference on Computer Vision and Pattern Recognition 1–12 (2008) 2378–2385.

[47] H.B. Li, F. Liu, Image denoising via sparse and redundant representations over learned dictionaries in wavelet domain, In: Proceedings of the Fifth International Conference on Image and Graphics (ICIG 2009), 2009, pp. 754–758.

[48] M. Protter, M. Elad, Image sequence denoising via sparse and redundant representations, IEEE Transactions on Image Processing 18 (2009) 27–35. (January).

[49] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Transactions on Image Processing 15 (2006) 3736–3745. (December).

[50] A. Dogandzic, K. Qiu, Automatic hard thresholding for sparse signal reconstruction from Nde measurements, Review of Progress in Quantitative Nondestructive Evaluation, Vols. 29a and 29b 1211 (2010) 806–813.

[51] I. Rish, G. Grabarnik, Sparse signal recovery with exponential-family noise, in: 2009 47th Annual Allerton Conference on Communication, Control, and Computing, Vols. 1 and 2, 2009, pp. 60–66.

[52] J.F. Murray, K. Kreutz-Delgado, Visual recognition and inference using dynamic overcomplete sparse learning, Neural Computation 19 (2007) 2301–2352. (September).

[53] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society Series B—Methodological 58 (1996) 267–288.

[54] I. Drori, D. Donoho, Solution of L1 minimization problems by LARS/homotopy methods, in: Proceedings of the 31th International Conference on Acoustics, Speech and Signal Processing, 2006, pp. 636–639.

[55] H.F. Li, T. Jiang, K.S. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Transactions on Neural Networks 17 (2006) 157–165. January).
[56] J. Liu, S.C. Chen, X.Y. Tan, D.Q. Zhang, Comments on efficient and robust feature extraction by maximum margin criterion, IEEE Transactions on Neural Networks 18 (2007) 1862–1864. (November).
[57] G. Stewart, Matrix Algorithms: Eigensystems: Society for Industrial and Applied Mathematics, University City Science Center, 2001.
[58] X.F. He, P. Niyogi, Locality preserving projections, Advances in Neural Information Processing Systems 16 16 (2004) 153–160.
[59] J. Yang, D. Zhang, J.Y. Yang, B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 650–664. (April).
[60] M. Wu, K. Yu, S. Yu, B. Sch lkopf, Local learning projections, in: Proceedings of the International Conference on Machine Learning (ICML), 2007, pp. 1039–1046.
[61] R. He, B.G. Hu, W.S. Zheng, X.W. Kong, Robust principal component analysis based on maximum correntropy criterion, IEEE Transactions on Image Processing, 20 (2011) 1485–1494.
[62] R. He, R. He, W.S. Zheng, B.G. Hu, Maximum correntropy criterion for robust face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 33 (2011) 1561–1576.

Jie Gui received the B.Sc. degree in Computer Science from Hohai University, Nanjing, China, 2004, the M.Sc. degree in Computer Applied technology from Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Science, Hefei, China, in 2007, and the Ph.D. degree in pattern recognition and intelligence system from University of Science and Technology of China, Hefei, China, in 2010. He is currently a postdoc in Institute of Automation, Chinese Academy of Sciences and he is also an assistant professor in Hefei Institute of Intelligent Machines, Chinese Academy of Science. His research interests are machine learning, pattern recognition, and image processing.

Zhenan Sun received the BE degree in industrial automation from Dalian University of Technology, the MS degree in system engineering from Huazhong University of Science and Technology, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 1999, 2002, and 2006, respectively. He is an associate professor at CASIA. In March 2006, he joined the Center of Biometrics and Security Research (CBSR) at the National Laboratory of Pattern Recognition (NLPR) of CASIA as a faculty. He is a member of the IEEE and the IEEE Computer Society. His current research focuses on biometrics, pattern recognition, and computer vision.

Wei Jia received the B.Sc. degree in informatics from Center of China Normal University, Wuhan, China, in 1998, the M.Sc. degree in computer science from Hefei University of technology, Hefei, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence system from University of Science and Technology of China, Hefei, China, in 2008. From June 2007 to February 2008, he was a research assistant in Biometrics Research Centre, Department of Computing, Hong Kong Polytechnic University. He is currently an associate professor in Hefei Institute of Intelligent Machines, Chinese Academy of Science. His research interests include palmprint recognition, pattern recognition, and image processing.

Rong-xiang Hu received B.Sc. degree in Computer science from Hefei University of Technology, Hefei, China, in 2006. From September 2006, he is a Master–Doctoral Program student in Department of Automation, University of Science and Technology of China, Hefei, China. His research interests include pattern recognition, machine learning and image processing.

Ying-Ke Lei received the BE degree in Communication Countermeasure Engineering from the Electronic Engineering Institute, China, in 1998, the Ph.D. degree in the Department of Automation, the University of Science and Technology of China. Now he is a lecturer in the Electronic Engineering Institute, Hefei, China. His research interests are machine learning, pattern recognition, and communication signal processing.

Shuiwang Ji received the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, in 2010. Currently, he is an Assistant Professor in the Department of Computer Science, Old Dominion University, Norfolk, VA. His research interests include machine learning, data mining, and bioinformatics. He received the Outstanding Ph.D. Student Award from Arizona State University in 2010.