Pattern Recognition Letters 32 (2011) 615–625

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Orthogonal local spline discriminant projection with application to face recognition

Ying-Ke Lei [a,b,c], Zhi-Guo Ding [c], Rong-Xiang Hu [a,b], Shan-Wen Zhang [a], Wei Jia [a,*]

[a] Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui 230031, China
[b] Department of Automation, University of Science and Technology of China, Hefei, Anhui 230027, China
[c] Electronic Engineering Institute, Hefei, Anhui 230037, China

ARTICLE INFO

Article history:
Received 8 April 2010
Available online 4 December 2010
Communciated by J. Yang

Keywords:
Feature extraction
Subspace learning
Manifold learning
Face recognition

ABSTRACT

In this paper, an efficient feature extraction algorithm called orthogonal local spline discriminant projection (O-LSDP) is proposed for face recognition. Derived from local spline embedding (LSE), O-LSDP not only inherits the advantages of LSE which uses local tangent space as a representation of the local geometry so as to preserve the local structure, but also makes full use of class information and orthogonal subspace to improve discriminant power. Extensive experiments on several standard face databases demonstrate the effectiveness of the proposed method.

1. Introduction

In the past two decades, appearance-based face recognition has attracted considerable interests in computer vision and pattern recognition. It is well known that there are two central issues in appearance-based face recognition: one is feature extraction for face representation; the other is classification of a new face image based on the extracted features. Generally, in many classical appearance-based methods, a face image of size $n_1 \times n_2$ pixels is represented as a vector in a $n_1 \times n_2$-dimensional space. Obviously, operating directly on such high-dimensional image space is ineffective and may lead to high computational and storage demands as well as poor performance. A typical way to circumvent the "curse of dimensionality" problem (Donoho, 2000) and other undesired properties of high-dimensional spaces is to use dimensionality reduction techniques.

Currently, many dimensionality reduction techniques for face recognition have been proposed, which can be broadly categorized into two classes: linear and nonlinear. Classical linear dimensionality reduction approaches, such as principal component analysis (PCA) (Jolliffe, 1989; Turk and Pentland, 1991) and linear discriminant analysis (LDA) (Duda et al., 2001), seek to find a meaningful low-dimensional subspace in a high-dimensional input space by linear transformation. This subspace can provide a compact representation of high-dimensional input data when the intrinsic structure of data embedded in the input space is linear. However, they may fail to discover the intrinsic structures of complex nonlinear data. In order to address this problem, a number of nonlinear dimensionality reduction techniques have been proposed. Among them, the manifold learning-based methods attracted extensive attention. The representative algorithms include isometric feature mapping (ISOMAP) (Tenenbaum et al., 2000; Silva and Tenenbaum, 2003; Law and Jain, 2006), locally linear embedding (LLE) (Roweis and Saul, 2000; Saul and Roweis, 2003), Laplacian eigenmaps (LE) (Belkin and Niyogi, 2003), Hessian-based locally linear embedding (HLLE) (Donoho and Grimes, 2003), maximum variance unfolding (MVU) (Weinberger and Saul, 2004), manifold charting (Brand, 2003), local tangent space alignment (LTSA) (Zhang and Zha, 2005), diffusion maps (Coifman and Lafon, 2006; Lafon and Lee, 2006), Riemannian manifold learning (RML) (Lin et al., 2006; Lin and Zha, 2008), and local spline embedding (LSE) (Xiang et al., 2006, 2009). Each manifold learning algorithm attempts to preserve a different geometrical property of the underlying manifold. Local approaches such as LLE, HLLE, LE, LTSA, RML and LSE aim to preserve the 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, some limitations are exposed when they are applied to pattern recognition.

One limitation is the out-of-sample problem. These nonlinear manifold learning algorithms yield maps that are defined only on the training data points, while how to evaluate the maps on novel test data points still attracts a lot of attention (He et al., 2005a,b). To overcome the drawback, Bengio et al. (2003), DeCoste (2001) proposed a kernel method to embed the new data points because of the generalization ability of Mercer kernel. He and Niyogi (2003) and He et al. (2005b) proposed a method named locality

* Corresponding author. Tel.: +86 0551 5591108.
E-mail address: icg.jiawei@gmail.com (W. Jia).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.11.024

preserving projection (LPP) to approximate the eigenfunctions of the Laplace–Beltrami operator on the manifold and the new testing points can be mapped to the learned subspace without trouble. He et al. (2005a), Zhang et al. (2007) introduced an explicit linear mapping to the original LLE and LTSA, respectively, which made it straightforward for handling new data samples. Yan et al. (2007) utilized the graph embedding framework for developing a novel algorithm called marginal Fisher analysis (MFA) to solve the out-of-sample problem. In addition, Chin and Suter (2008) tuned data-dependent kernel functions derived from Gaussian basis functions for extrapolating manifolds learned via MVU to novel out-of-sample data. Among the approaches mentioned above, the linear approaches based on manifold learning can address the out-of-sample problem with the cheapest computational cost.

Another limitation is that classical manifold learning approaches neglect the class information, which will inevitably lead to a heavy weakening of their performances on pattern recognition. Many modified manifold learning-based approaches have been recently proposed to make use of the label information. A typical approach to overcome this limitation of manifold learning for pattern recognition is to modify input space distances by taking into account class labels of individual data points, such as supervised ISOMAP (Geng et al., 2005) using a certain kind of dissimilarity based on Euclidean distance, supervised LLE (Ridder et al., 2003) using Euclidean distance considering the known class label information, probability-based LLE (Zhang and Zhao, 2007) using a probability-based distance, and weighted locally linear embedding (Pan et al., 2008) using a cam weighted distance. These supervised manifold learning approaches have achieved good classification performance on some data sets. At the same time, they tend to divide all the sample points into disconnected parts instead of an entire neighborhood graph, which also bring a problem about how to apply original manifold learning approaches to disconnected components. Alternatively, some approaches combine original manifold learning techniques with supervised linear subspace methods. The seminal approach is LLE + LDA (Zhang et al., 2004), which comprises two steps. Sample points are first mapped into the intrinsic low-dimensional space based on LLE and then LDA is adopted to enhance between-class distances and decrease within-class distances. There are still some weaknesses in this proposed approach. Firstly, the embedding dimension of LLE must be reduced to be smaller than the number of the classes in order to avoid the small sample size problem (SSS), thus some useful classification information may be discarded. Furthermore, the simple addition of LLE and LDA makes the two-step approach more complicated. Another alternative is to add a penalty term to the cost function favoring embeddings with small within-class distances (Pang et al., 2006). Such approaches must manage a tradeoff between class discrimination and trustworthiness of the visualizations.

In this paper, inspired by the idea of LSE (Xiang et al., 2006, 2009), we propose a novel linear subspace learning technique, called orthogonal local spline discriminant projection (O-LSDP). The tangent space in the neighborhood of each data point is firstly built in our method which can represent the local geometry of the intrinsic manifold structure. According to the notion of the compatible mapping in LSE, smooth splines are constructed to align those local tangent spaces to its own single low-dimensional global coordinates. We then compute a transformation matrix which maps the data points to a subspace. The linear transformation matrix is obtained by optimizing an objective function, which captures the discrepancy of the local geometries in the reduced space and introduces the maximum margin criterion (MMC) (Li et al., 2006) simultaneously. Therefore, our method effectively combines the ideas of LSE and MMC, i.e. it can hold the strong discriminant power of MMC and preserve the intrinsic geometry of the data samples simultaneously. In order to improve the discriminant power, we present a new method for obtaining a set of orthogonal basis eigenvectors.

O-LSDP constructs a local tangent space at each data point, which models explicitly the data topology. Similar to LSE, the corresponding tangent space projection is estimated to capture the geometry of the neighborhood of each point and those local tangent coordinates are nonlinearly aligned in the reduced space by different spline functions to obtain a global coordinate system. However, in contrast to LSE, O-LSDP has a number of desirable properties:

1. O-LSDP computes an explicit linear mapping from the input space to the reduced space. Note that in LSE, the mapping is implicit and it is not clear how new data samples can be embedded.
2. O-LSDP attempts to manage the trade-off between MMC, which emphasizes discriminant power, and LSE, which is based mainly on preserving local structure.
3. O-LSDP seeks to find a set of orthogonal basis functions and significantly improves its recognition accuracy.

The rest of this paper is organized as follows: Section 2 describes the MMC and LSE algorithms. The O-LSDP algorithm is developed in Section 3. Section 4 demonstrates the experimental results. Finally, conclusions are presented in Section 5.

2. Related works

Given a data set of n data points $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m\times n}$, the goal of dimensionality reduction is to project the high-dimensional data into a low-dimensional feature space. Let us denote the corresponding set of n points in the reduced space as $Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{d\times n}$, with $d \ll m$, in which y_i is a low-dimensional representation of x_i ($i = 1, 2, \ldots, n$).

2.1. Maximum margin criterion

MMC (Li et al., 2006) aims at maximizing the average margin between classes in the projected space. Let S_w and S_b be the within-class scatter matrix and the between-class scatter matrix defined by

$$S_w = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_j^i - m_i)(x_j^i - m_i)^T, \tag{1}$$

$$S_b = \sum_{i=1}^{c} n_i(m_i - m)(m_i - m)^T, \tag{2}$$

where c is the number of classes, m is the total sample mean vector, m_i is the average vector of the ith class, n_i is the number of samples in the ith class, and x_j^i is the jth sample in the ith class. The objective function of MMC under projection matrix W is

$$J(W) = tr\{W^T(S_b - S_w)W\}. \tag{3}$$

Confining the column vectors in W to be unit vectors, W that maximizes Eq. (3) can be calculated through the following eigenvalue equation

$$(S_b - S_w)w = \lambda w. \tag{4}$$

Comparing MMC with the classical LDA, we find that the former avoids calculating the inverse within-class scatter, i.e. $(S_w)^{-1}S_b$ is substituted by $S_b - S_w$. This can not only make the computation more efficient but also avoid the SSS problem of the within-class scatter.

2.2. Local spline embedding

LSE (Xiang et al., 2006, 2009) is a recently proposed manifold learning method for nonlinear dimensionality reduction. This method is developed from the framework of part optimization and whole alignment. Each data point is represented in different local coordinate systems by part optimization. But its global coordinate should be maintained unique. Whole spline alignment is used to achieve this goal. The outline of LSE can be summarized as follows:

Step 1: Identify neighbors. For each data point x_i, Let $X_i = [x_{i_1}, x_{i_2}, \ldots, x_{i_k}] \in \mathbb{R}^{m\times k}$denote the collection of its k nearest neighbors. Use the *KNN* or ε-ball criterion to identify the indices corresponding to the k nearest neighbors.

Step 2: Obtain tangent coordinates. Perform a singular value decomposition of the centralized matrix of X_i, we have

$$X_i H = U_i \sum_i V_i^T, \quad i = 1, \ldots, n,$$

where H is the centering operator. The local tangent space coordinates can be obtained from the following formula:

$$\Theta_i = U_i^T X_i H = [\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i)}], \tag{5}$$

where $\theta_j^{(i)}$ is the local tangent coordinate of the jth nearest neighbor of data point x_i. In essence, this step is equal to performing a local principal component analysis (PCA); Θ_i is the projection of the points in a local neighborhood on the local PCA.

Step 3: Align global coordinates. For the ith local tangent space projection Θ_i, let $Y_i = [y_{i_1}, y_{i_2}, \ldots, y_{i_k}] \in \mathbb{R}^{d\times k}$ contain the corresponding global coordinates of the k data points in Θ_i. Further, denote the rth row of Y_i by $[y_{i_1}^{(r)}, y_{i_2}^{(r)}, \ldots, y_{i_k}^{(r)}]$. We determine d spline functions $g_i^{(r)} : \mathbb{R}^d \mapsto \mathbb{R}, r = 1, 2, \ldots, d$, such that the coordinate components can be faithfully mapped:

$$y_{i_j}^{(r)} = g_i^{(r)}(\theta_j^{(i)}), \quad j = 1, 2, \ldots, k. \tag{6}$$

Because Y_i is unknown, the desirable splines not only can satisfy the conditions in Eq. (6) but also make the reconstruction error to be formulated explicitly in terms of Y_i. The spline developed in Sobolev space meets our tasks:

$$g^{(r)}(t) = \sum_{i=1}^{l} \beta_i^r p_i(t) + \sum_{j=1}^{k} \alpha_j^r \phi_j(t), \quad r = 1, 2, \ldots, d, \tag{7}$$

where $\{p_i(t)\}_{i=1}^{l}$ are a set of polynomials in $\mathbb{R}^d$, and φ_j is a Green's function (Xiang et al., 2009). Let

$$A_i = \begin{pmatrix} K & P \\ P^T & 0 \end{pmatrix} \in \mathbb{R}^{(k+l)\times(k+l)}, \tag{8}$$

where K is $k \times k$ symmetrical matrix with elements $K_{st} = \phi(\|\theta_s^{(i)} - \theta_t^{(i)}\|)$, and P is a $k \times l$ matrix with elements $P_{ts} = p_s(\theta_t^{(i)})$. Then, The coefficients $\alpha^r = [\alpha_1^r, \alpha_2^r, \ldots, \alpha_k^r]^T \in \mathbb{R}^k$ and $\beta^r = [\beta_1^r, \beta_2^r, \ldots, \beta_l^r]^T \in \mathbb{R}^l$ in Eq. (7) can be solved via the following linear equations:

$$A_i \cdot \begin{pmatrix} \alpha^1 & \ldots, & \alpha^d \\ \beta^1 & \ldots, & \beta^d \end{pmatrix} = \begin{pmatrix} Y_i^T \\ 0 \end{pmatrix}. \tag{9}$$

To preserve as much of the local geometry in the low-dimensional feature space, we intend to find Y_i to minimize the penalized reconstruction error, i.e.

$$E(Y_i) = \sum_{r=1}^{d} \sum_{j=1}^{k} (y_{i_j}^{(r)} - g_i^{(r)}(\theta_j^{(i)}))) + \lambda \sum_{r=1}^{d} (\alpha^r)^T K \alpha^r. \tag{10}$$

Here, the regularization parameter λ controls the amount of smoothness of the spline. With an enough small λ, the first term on the right of Eq. (10) can be neglected. Therefore, we have

$$E(Y_i) \propto \sum_{r=1}^{d} (\alpha^r)^T K \alpha^r = tr(Y_i B_i Y_i^T), \tag{11}$$

where B_i is the upper left $k \times k$ subblock of A_i^{-1}.

Summing all the reconstruction errors together, we have

$$E(Y) = \sum_{i=1}^{n} tr(Y_i B_i Y_i^T). \tag{12}$$

Let S_i be a column selection vector such that $YS_i = Y_i$. The objective function is converted to the following form:

$$E(Y) = tr(YSBS^T Y^T) = tr(YMY^T), \tag{13}$$

where $S = [S_1, \ldots, S_n]$, $B = diag(B_1, \ldots, B_n)$, and $M = SBS^T$.

To uniquely determine Y, we impose the constraint $YY^T = I$. Then, the minimum of $E(Y)$ for the d-dimensional global embedding is given by the d eigenvectors of the matrix M, corresponding to the 2nd to $(d+1)$st smallest eigenvalues of M.

3. Orthogonal local spline discriminant projection

In this section, we propose a new linear subspace algorithm based on LSE and MMC. Firstly, the linearization of the original LSE is presented. Then we propose a new method which makes class separability and neighborhood structure preservation to be attained at the same time. In order to improve the discriminability of the proposed method, we also present a method for obtaining orthogonal basis functions which can provide a more faithful representation for the input data.

3.1. A linear approximation to the original LSE

It is well known that the original LSE algorithm might be unsuitable for pattern recognition tasks because it yield an embedding only based on the training data set. In order to overcome the out-of-sample problem, an explicit linear mapping from X to Y, i.e. $Y = V^T X$, is imposed. Thus the objective function for the original LSE can be converted to the following form:

$$J_1(Y) = \min tr(YMY^T) = \min tr(V^T XMX^T V). \tag{14}$$

Once linear transformation matrix V is determined, mapping new data points to the lower dimensional space becomes trivial. We refer to this algorithm as the linearization of LSE (LLSE). Considering a new test data sample x_t that needs to be mapped, the test sample is projected onto the subspace using the dimensionality reduction matrix V. So we have $y_t = V^T x_t$ and mapping the new data point reduces to a simple matrix vector product. Fig. 1a shows a simple artificial example that LLSE successfully solves the two-class classification problem. In this example, the data of the first class is represented as a single Gaussian distribution, while the second is represented as two separated Gaussians. The dash-dot line and dashed one represent the learned optimal projection directions from LLSE and MMC, respectively. The result indicates that LLSE can effectively find an optimal projection such that the multimodal data samples can preserve their intuitively natural way in the reduced space whereas MMC collapses samples of different classes into a single cluster. Moreover, the experimental results in Section 4 also demonstrate that LLSE can achieve satisfactory performances in some real world data sets. But we have to confess that LLSE does not take class information into account, and so the linear transformation derived from LLSE is not always optimal for recognition problem. For example, Fig. 1b shows another

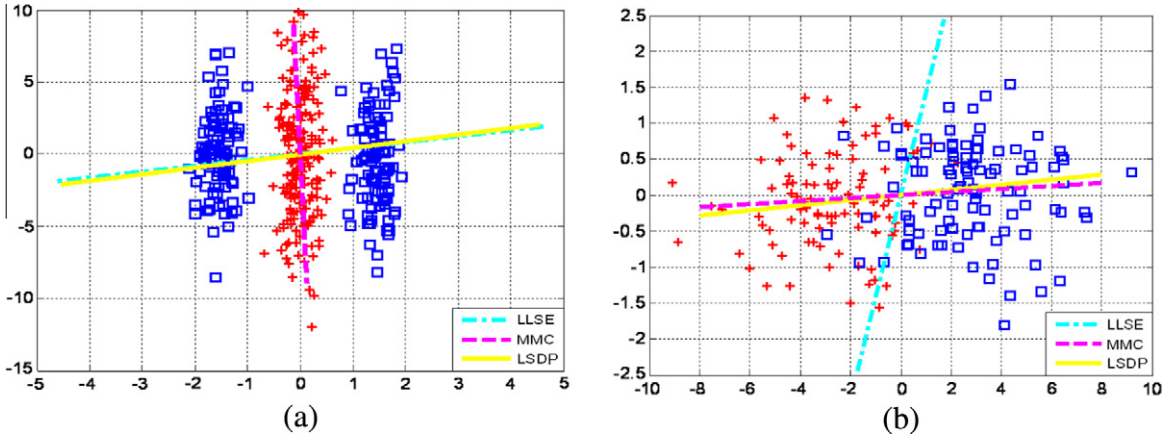

Fig. 1. Examples of dimensionality reduction by LLSE, MMC and LSDP. (a) Toy data set 1. (b) Toy data set 2.

example where 300 sample points from the two classes are mapped into 1-dimensional feature subspace by LLSE and MMC. In this artificial problem, two classes are represented as two different Gaussian distributions. Obviously, two data clusters are so close that there are some overlapping data points between two classes. It can be seen that LLSE mixes the samples of different classes into one cluster and fail to find the optimal direction due to its unsupervised nature. In contrast to LLSE, MMC can find a more discriminative direction.

3.2. Optimal linear discriminant projection

Based on the analysis mentioned above, it can be found that the linear approximation to the original LSE (LLSE) seeks to preserve as much as possible local structure defined by the nearest neighbors. It often fails to preserve within-class local geometry, which is very important for pattern classification, because the nearest neighbors may belong to different classes due to influence of complex variations, such as pose, illumination, and expression. Thus, in order to obtain optimal linear discriminant projection, we introduce the MMC presented above to the LLSE such that it can preserve the intrinsic geometry of the neighbors as LLSE (Xiang et al., 2009; He et al., 2005a,b; Kokiopoulou and Saad, 2007) and at the same time hold the strong discriminant power of MMC. That is to say, the linear transformation obtained by LLSE can satisfy Eq. (3) simultaneously. Then, the problem can be written as the following multi-object optimization:

$$\begin{cases} \min tr\{V^T X M X^T V\}, \\ \max tr\{V^T (S_b - S_w) V\}. \end{cases} \tag{15}$$

Furthermore, there are two constraints in LSE, that is

$$V^T X X^T V = I, \tag{16}$$

$$Ye = 0. \tag{17}$$

Eq. (17) requires the outputs $\{y_i\}_{i=1}^n$ to be centered on the origin, which removes the translational degree of freedom. However, in order to improve the discriminant ability of our proposed method, MMC is chosen as a criterion to implicitly construct different optimal translation and rescaling operators for each class. Therefore, in contrast to the original LSE, our proposed method should neglect this centering constraint for maximizing the average margin between different classes in the embedded space. Then, the above optimization can be deduced to solve the following constrained objective function:

$$\{ \min tr\{V^T X M X^T V\} max tr\{V^T (S_b - S_w) V\}$$

$$\text{s.t.} \quad V^T X X^T V = I. \tag{18}$$

The constrained multi-object optimization is conducted to minimize the reconstruction error and maximize the margin between difference classes simultaneously. We formulate this discriminator by using the linear manipulation as follows:

$$\min tr\{V^T (X M X^T - (S_b - S_w)) V\}$$

$$\text{s.t.} \quad V^T X X^T V = I. \tag{19}$$

It is easily shown that the above optimization problem can be converted into solving a generalized eigenvalue problem as follows:

$$(X M X^T - (S_b - S_w)) v = \lambda X X^T v. \tag{20}$$

Let the column vectors $v_1, v_2, \ldots, v_d$ be the d smallest generalized eigenvectors of $XMX^T - (S_b - S_w)$ and XX^T corresponding to the d smallest eigenvalues. The transformation matrix V which minimizes the objective function is as follows:

$$V = [v_1, v_2, \ldots, v_d]. \tag{21}$$

In practical problems, one often suffers from the difficulty that XX^T is singular. This stems from the fact that sometimes the number of samples in the training set is much smaller than the dimension of each data point. To address the complication of a singular XX^T, a PCA step is adopted to project the data set to a PCA subspace so that the resulting matrix XX^T is nonsingular.

We call the new linear subspace method as local spline discriminant projection (LSDP). With using LSDP, we perform the classification task on the two toy data sets shown in Fig. 1. The results demonstrate that LSDP holds the advantages of both MMC and LLSE.

3.3. Obtaining orthogonal eigenvectors

The generalized eigenvectors obtained by solving Eq. (20) are nonorthogonal. This makes it difficult to obtain the optimal low-dimensional representation of the original high-dimensional input data. In this paper, we propose a method for producing orthogonal basis functions, which is called O-LSDP. Note that the derivation presented here is motivated by (Duchene and Leclercq, 1988).

Let $L = XMX^T - (S_b - S_w)$. The O-LSDP algorithm seeks to find a set of orthogonal basis vectors $v_1, v_2, \ldots, v_d$ by solving the following optimization problem:

$$\min tr\{V^T L V\}$$

$$\text{s.t.} \quad v_1^T v_2 = v_1^T v_3 = \cdots = v_{d-1}^T v_d = 0,$$
$$v_1^T X X^T v_1 = v_2^T X X^T v_2 = \cdots = v_d^T X X^T v_d = 1. \tag{22}$$

It is easy to check that v_1 is the eigenvector of the generalized eigenproblem

$$Lv = \lambda XX^T v$$

associated with the smallest eigenvalue. Since XX^T is always nonsingular in the PCA subspace, v_1 is the eigenvector of the matrix $(XX^T)^{-1}L$ associated with the smallest eigenvalue.

In order to get the kth basis vector, we minimize the following objective function

$$\min\{v_k^T L v_k\} \tag{23}$$

with the constraints

$$v_1^T v_k = v_2^T v_k = \cdots = v_{k-1}^T v_k = 0, \quad v_k^T XX^T v_k = 1.$$

To solve the above optimization problem, we use the Lagrangian multiplier:

$$J_k = v_k^T L v_k - \lambda(v_k^T XX^T v_k - 1) - \mu_1 v_1^T v_k - \cdots - \mu_{k-1} v_{k-1}^T v_k.$$

We set the partial derivative of J_k with respect to v_k to zero and obtain

$$2Lv_k - 2\lambda XX^T v_k - \mu_1 v_1 - \cdots - \mu_{k-1} v_{k-1} = 0. \tag{24}$$

Multiplying the left side of Eq. (24) by v_k^T, we obtain

$$2v_k^T L v_k - 2\lambda v_k^T XX^T v_k = 0. \tag{25}$$

Multiplying the left side of Eq. (24) successively by $v_1^T(XX^T)^{-1}, \ldots, v_{k-1}^T(XX^T)^{-1}$, now we can obtain a set of k-1 equations as follows:

$$\begin{aligned}
&\mu_1 v_1^T(XX^T)^{-1} v_1 + \cdots + \mu_{k-1} v_1^T(XX^T)^{-1} v_{k-1} = 2v_1^T(XX^T)^{-1} L v_k,\\
&\mu_1 v_2^T(XX^T)^{-1} v_1 + \cdots + \mu_{k-1} v_2^T(XX^T)^{-1} v_{k-1} = 2v_2^T(XX^T)^{-1} L v_k,\\
&\qquad\qquad \ldots\ldots\\
&\mu_1 v_{k-1}^T(XX^T)^{-1} v_1 + \ldots + \mu_{k-1} v_{k-1}^T(XX^T)^{-1} v_{k-1} = 2v_{k-1}^T(XX^T)^{-1} L v_k.
\end{aligned} \tag{26}$$

We define

$$\boldsymbol{\mu}_{k-1} = [\mu_1, \ldots, \mu_{k-1}]^T, \boldsymbol{V}_{k-1} = [v_1, \ldots, v_{k-1}], \quad \text{and}$$
$$\boldsymbol{Q}_{k-1} = V_{k-1}^T (XX^T)^{-1} \boldsymbol{V}_{k-1}.$$

So Eq. (26) can be represented in a matrix equation

$$\boldsymbol{Q}_{k-1}\boldsymbol{\mu}_{k-1} = 2\boldsymbol{V}_{k-1}^T(XX^T)^{-1} L v_k.$$

Thus

$$\boldsymbol{\mu}_{k-1} = 2\boldsymbol{Q}_{k-1}^{-1}\boldsymbol{V}_{k-1}^T(XX^T)^{-1} L v_k. \tag{27}$$

Let us now multiply the left side of Eq. (24) by $(XX^T)^{-1}$

$$2(XX^T)^{-1} L v_k - 2\lambda v_k - \mu_1 (XX^T)^{-1} v_1 - \cdots - \mu_{k-1}(XX^T)^{-1} v_{k-1} = 0.$$

This can be expressed using matrix notation as

$$2(XX^T)^{-1} L v_k - 2\lambda v_k - (XX^T)^{-1}\boldsymbol{V}_{k-1}\boldsymbol{\mu}_{k-1} = 0.$$

With Eq. (27), we obtain

$$\{\boldsymbol{I} - (XX^T)^{-1}\boldsymbol{V}_{k-1}\boldsymbol{Q}_{k-1}^{-1}\boldsymbol{V}_{k-1}^T\}(XX^T)^{-1} L v_k = \lambda v_k. \tag{28}$$

As shown in Eq. (25), λ is just the criterion to be minimized, thus v_k is the eigenvector of

$$\boldsymbol{R}_k = \{\boldsymbol{I} - (XX^T)^{-1}\boldsymbol{V}_{k-1}\boldsymbol{Q}_{k-1}^{-1}\boldsymbol{V}_{k-1}^T\}(XX^T)^{-1} L \tag{29}$$

associated with the smallest eigenvalue of $\boldsymbol{R}_k$.

According to the above preparation, the main steps for our O-LSDP algorithm can be summarized as follows:

Step 1: Project the data set X into the PCA subspace by discarding the minor components.

Step 2: For each data point x_i, determine its k nearest neighbors by *KNN* or ε-ball algorithm.

Step 3: Compute the d left singular vector matrix U_i of X_iH. Set $\boldsymbol{\Theta}_i$ as in Eq. (5).

Step 4: Compute matrix A_i based on Eq. (8).

Step 5: Construct spline alignment matrix M by locally summing as follows:

$$M(I_i, I_i) \leftarrow M(I_i, I_i) + B_i, \quad i = 1, 2, \ldots, n$$

with the initial $M = 0$, where $I_i = \{i_1, \ldots, i_k\}$ denotes the set of indices for the k nearest neighbors of x_i and B_i is the upper left $k \times k$ subblock of A_i^{-1}.

Step 6: Compute matrix XMX^T.

Step 7: Compute the between-class scatter S_b, within-class scatter S_w, and their difference $S_b - S_w$, respectively.

Step 8: Compute the d orthogonal basis vectors $V = [v_1, v_2, \ldots, v_d]$ based on Eq. (28) and obtain the d dimensional projection $Y = V^T X$.

4. Experimental results

This section evaluates the performance of the proposed O-LSDP in comparison with seven representative algorithms, i.e., MMC (Li et al., 2006), LDA (Belhumeour et al., 1997), SLPP (Cai et al., 2005), supervised LLTSA (SLLTSA) (Zhang et al., 2007), marginal Fisher analysis (MFA) (Yan et al., 2007), the linearization of LSE (LLSE), and local spline discriminant projection (LSDP), on three face image databases, i.e., Yale database,[1] Olivetti Research Laboratory (ORL) database,[2] and Extended Yale B database.[3] Among these algorithms, SLPP, SLLTSA, MFA, LLSE, and LSDP are manifold learning-based algorithms. Preprocessing was performed to crop all face images from three databases. The size of each cropped image in all the experiments is 32 × 32 pixels, with 256 gray levels per pixel. Thus, each image can be represented by a 1024-dimensional vector in an image space. No further preprocessing is done.

The k nearest neighborhood parameter for constructing the nearest neighbor graph in SLPP, SLLTSA, LLSE, LSDP, and O-LSDP can be chosen as $k = l - 1$, where l denotes the number of training samples per class. The justification for this choice is that the l samples of the same class should be located in the same local geometrical structure provided that within-class samples are well clustered in the observation space. For the MFA method, the important parameters include k_1 (the number of the nearest in-class neighbors) and k_2 (the number of the closest out-of-class sample pairs). We chose the best k_1 between one and $l - 1$. We similarly selected the best k_2 between l and $8c$, where c denotes the number of classes. Note that all algorithms involve a PCA phase. In this phase, we kept 100% image energy and selected all principal components corresponding to the non-zero eigenvalues for each method. Different pattern classifiers have been applied for face recognition, including KNN, Support Vector Machine, and Neural Network (Huang, 1996, 1998, 1999; Huang and Ma, 1999), etc. In this study, we adopt the 1-NN classifier for its simplicity. The Euclidean metric is used as our distance measure.

4.1. Yale database

The Yale database contains 165 gray-scale images of 15 individuals (each person providing 11 different images). The images demonstrate variations in lighting condition (left-light, center-light, right-light), facial expression (normal, happy, sad, sleepy, surprised,

[1] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.
[2] http://www.cam-orl.co.uk/facedatabase.html.
[3] http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html.

and wink), and with/without glasses. Fig. 2 shows sample images of one person.

Firstly, we test the impact of selecting different dimensions in the reduced subspace on the recognition rate. During the testing phases, the 1*NN* classifier was used. Note that, for LDA, there are at most $c-1$ nonzero generalized eigenvalues, and so an upper bound on the dimension of the reduced space is $c-1$. Fig. 3 illustrates the recognition rates versus the variation of subspace dimensions when 3, 5, and 7 images per individual were randomly selected for training. In general, the performance of all these methods varies with the number of dimensions. At the beginning, the recognition rates improve with the increase of the dimensions. However, more dimensions will not lead to higher recognition rate after these methods attain the best results.

Secondly, we randomly select the seven images as training sets and the rest four images as testing sets for each class. The training sets were used to learn the low-dimensional subspace with the projection matrix. The testing sets were used to report the final recognition accuracy. Fig. 4 shows the best mean recognition rates for 20 times. It can be found that our proposed method outperforms the other techniques. The recognition approaches the maximal average results at 77.83(±5.02)%, 82.25(±4.72)%, 82.92(±4.71)%, 82.25(±5.52)%, 83.42(±4.03)%, 78.42(±4.24)%, 82.33(±4.73)%, and 85.50(±3.29)% for MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP, respectively.

Thirdly, the experiments are conducted to examine the effect of the training number on the performance. For each method, five random subsets with three, four, five, six, seven images per individual were selected for training. The rest of the database was used for testing. For the baseline method, the recognition is simply performed in the original 1024-dimensional image space without any dimensionality reduction. Such a trail was independently performed 20 times, and then the average recognition results were calculated. Table 1 shows the maximal average recognition accuracy,

Fig. 2. Sample images from one person in the Yale database.

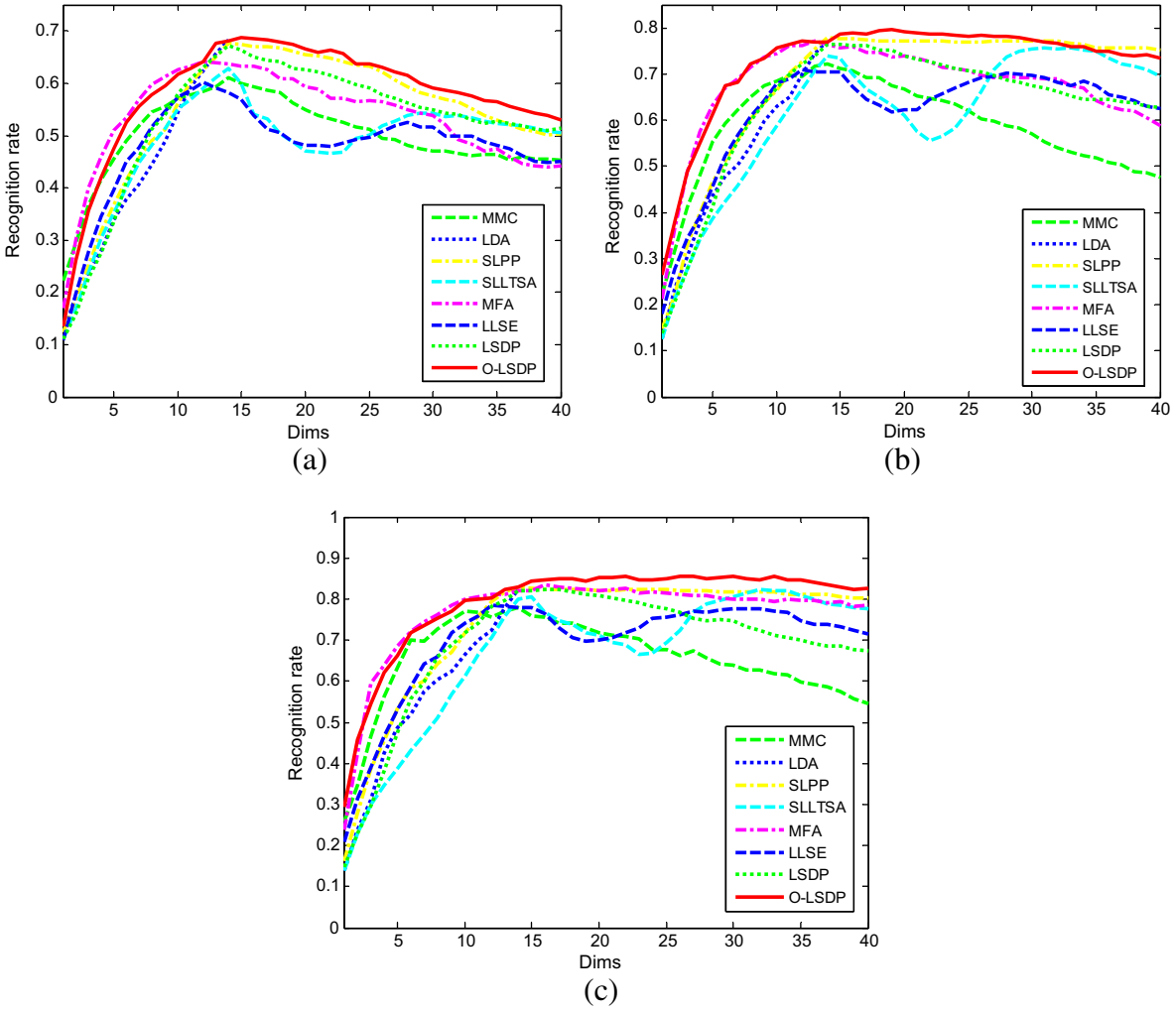

Fig. 3. The recognition rates of MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP versus the dimensions on the Yale database. (a) Three samples for training. (b) Five samples for training. (c) Seven samples for training.

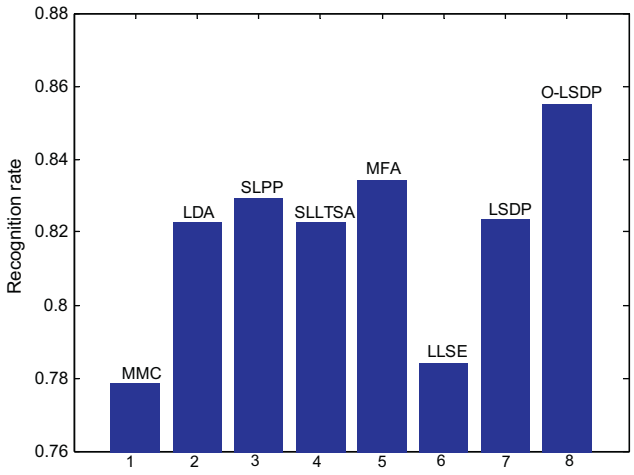

Fig. 4. Performance comparison of best mean recognition rates using MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP on the Yale database.

the corresponding standard deviations (*std*), and the reduced dimensions for MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP. As can be seen, O-LSDP significantly outperforms the other algorithms among all the cases.

4.2. ORL database

The ORL face database contains 400 face images of 40 individuals (each one has ten images). The images were captured at different times and have different variations including expressions (open or closed eyes, smiling or nonsmiling) and facial details (glasses or no glasses). The images were taken with a tolerance for some tilting and rotation of the face up to 20°. Ten sample images of one individual are displayed in Fig. 5.

The experimental design is the same as before. We averaged the results over 20 random splits. The recognition rates versus the variation of dimensions with 3, 5, and 7 images per individual randomly selected for training are illustrated in Fig. 6. From Fig. 6, we can see that the discrimination power of these methods will be enhanced with the increase of final projected dimensions, but they will not increase all the time. When the final dimensions are higher than some threshold, the final recognition rates will stand still.

Fig. 7 shows the best mean recognition rates for 20 times, where seven images for each individual were randomly selected for training and the rest were used for testing. As can be seen, O-LSDP algorithm significantly performs the best, while SLLTSA performs poorly. Besides, LDA and SLPP almost achieve the same accuracy rate. We investigate the maximal average recognition accuracy at 36, 39, 39, 111, 51, 30, 39, and 34 dimensions for MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP, respectively. The best mean recognition rates of MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP are 94.92%, 95.71%, 95.79%, 94.21%, 96.46%, 94.04%, 95.42%, and 97.08%, and the standard deviations are 1.77%, 1.86%, 1.90%, 1.92%, 1.97%, 2.50%, 2.00% and 2.12%, respectively. The corresponding face subspaces obtained by carrying out the methods mentioned above are called optimal face subspace for each method.

Moreover, the effect of the training sample number is also tested in the following experiment. We randomly selected 3, 4, 5, 6, and 7 training samples and then the rest samples for test ones. We repeated these trails 20 times and computed the average results. The best result obtained in the optimal subspace and their corresponding standard deviations and dimensions for each method are shown in Table 2. It can be seen that our O-LSDP algorithm significantly performs the best among all the cases. LLSE yields the lowest recognition rate. MMC performs better than SLLTSA, and achieves comparable performance to LDA with the increase of the number of training samples. LDA, SLPP, and LSDP performed comparably to each other. The performance of MFA approaches that of our algorithm as the number of training samples is increased.

4.3. Extended Yale B database

The Extended Yale B database consists of 2414 frontal-face images of 38 individuals, where each subject has about 64 images captured under various laboratory-controlled lighting conditions. We simply used the cropped images and resized them to 32×32 pixels. Sixty-four sample images of one individual are displayed in Fig. 8. It is well known that MFA improves its discriminant ability at the cost of affording intolerable computational complexity. Therefore, in this experiment, we did not use the MFA method for comparison. For training, we randomly selected different numbers (10, 30, 50) of images per individual, and used the rest images for testing. Such a trial was independently performed 20 times, and then the average recognition results were calculated. Fig. 9 shows the average recognition rates versus subspace dimensions. The

Table 1
The maximal average recognition rates and the corresponding standard deviations (%) with the reduced dimensions for BASELINE, MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP on the Yale database.

Method	3 Train	4 Train	5 Train	6 Train	7 Train
Baseline	50.79 ± 4.53	54.19 ± 4.94	56.00 ± 5.44	59.20 ± 4.61	60.58 ± 4.87
MMC	61.08 ± 4.84 (14)	67.43 ± 5.20 (13)	72.11 ± 3.42 (14)	75.40 ± 4.29 (14)	77.83 ± 5.02 (14)
LDA	68.25 ± 4.21 (14)	74.86 ± 5.51 (14)	77.22 ± 3.50 (14)	81.73 ± 5.05 (14)	82.25 ± 4.72 (14)
SLPP	67.92 ± 4.25 (14)	75.14 ± 5.46 (16)	77.22 ± 3.50 (14)	81.60 ± 4.94 (14)	82.92 ± 4.71 (14)
SLLTSA	62.75 ± 4.30 (14)	69.71 ± 5.24 (15)	75.72 ± 4.62 (33)	80.27 ± 4.69 (32)	82.25 ± 5.52 (32)
MFA	64.04 ± 5.63 (12, 2, 117)	71.48 ± 5.82 (12, 3, 115)	77.06 ± 4.19 (13, 4, 77)	80.80 ± 4.82 (17, 5, 99)	83.42 ± 4.03 (16, 4, 109)
LLSE	60.25 ± 4.54 (12)	65.14 ± 5.15 (13)	70.89 ± 5.77 (12)	76.00 ± 3.82 (13)	78.42 ± 4.24 (12)
LSDP	67.17 ± 5.05 (14)	73.67 ± 5.82 (14)	76.50 ± 3.52 (15)	80.93 ± 4.70 (15)	82.33 ± 4.73 (16)
O-LSDP	**68.75 ± 4.76 (15)**	**76.48 ± 4.43 (16)**	**79.72 ± 3.69 (19)**	**84.27 ± 3.95 (22)**	**85.50 ± 3.29 (26)**

For MFA, the first numbers in the parentheses are the selected subspace dimensions, the second and the third numbers are the parameters k_1 and k_2, respectively.

Fig. 5. Sample images from one person in the ORL database.

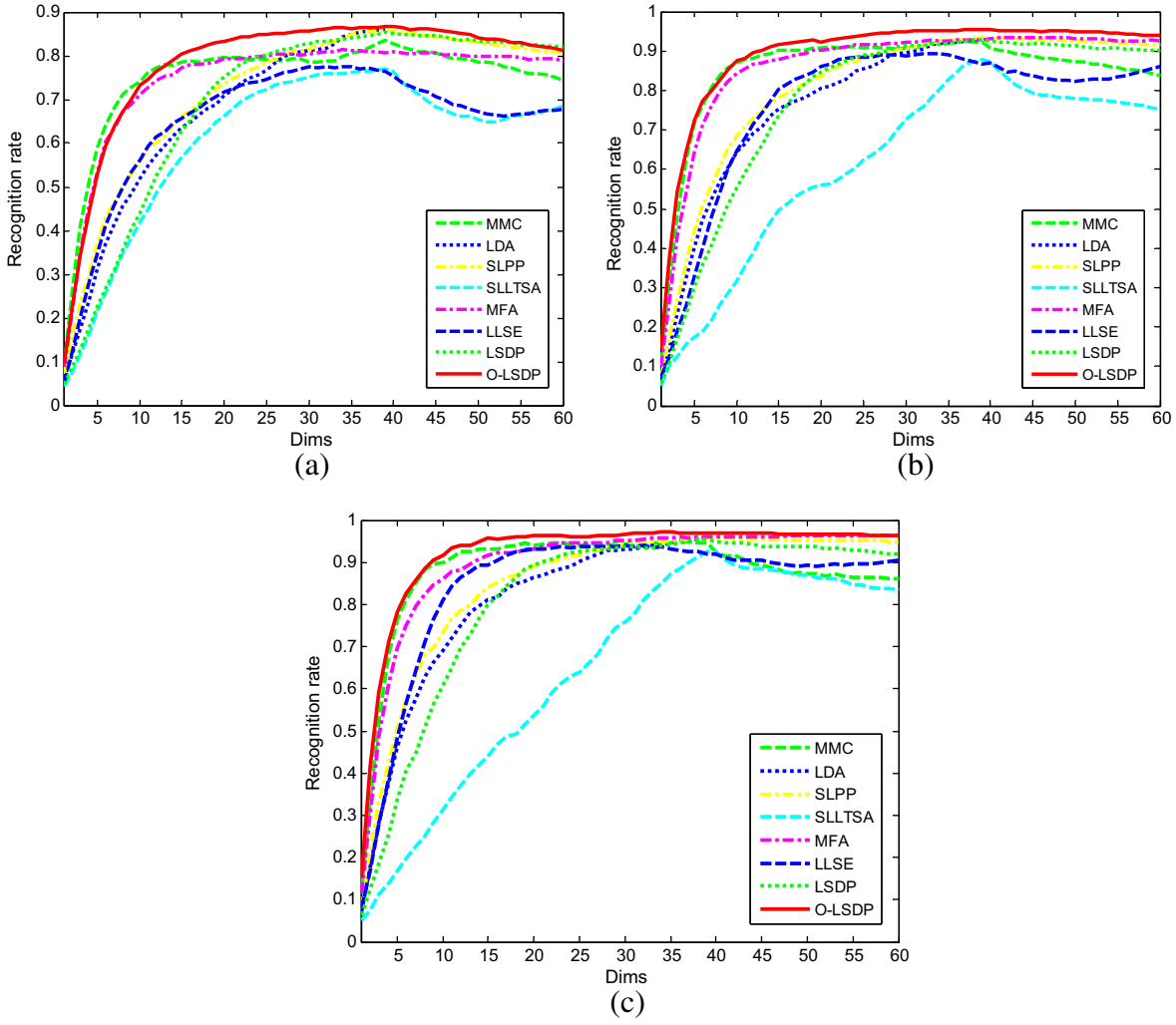

Fig. 6. The recognition rates of MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP versus the dimensions on the ORL database. (a) Three samples for training. (b) Five samples for training. (c) Seven samples for training.

maximal recognition rate of each method and the corresponding standard deviation with the reduced dimension are given in Table 3. From Table 3, we can see two main points. First, O-LSDP outperforms MMC, LDA, SLPP, SLLTSA, LLSE, and LSDP, whether the training sample size is 10, 30, or 50. Second, the orthogonalization step can significantly improve the performance of LSDP.

4.4. Discussion

Several experiments have been conducted on three different face databases. Here, it is necessary to highlight some observations about these tests:

1. The proposed O-LSDP consistently achieves the best recognition rate in all the experimental cases. The data sets used in this study are Yale, ORL, and Extended Yale B face databases. The images for each person vary from pose, illumination to expression. Some research efforts have shown that such face datasets may reside on or close to a low-dimensional sub-manifold embedded in the ambient space (Tenenbaum et al., 2000; Silva and Tenenbaum, 2003; Roweis and Saul, 2000; Saul and Roweis, 2003). Different from PCA and LDA which see only the Euclidean structure of face space, our proposed method explicitly considers the face manifold structure which is modeled by a neighborhood graph. Moreover, our proposed method is obtained by finding the optimal linear approximations to the original nonlinear local spline embedding. Therefore, it can efficiently extract intrinsic features that preserve local information, and obtain a face subspace that best detects the essential face

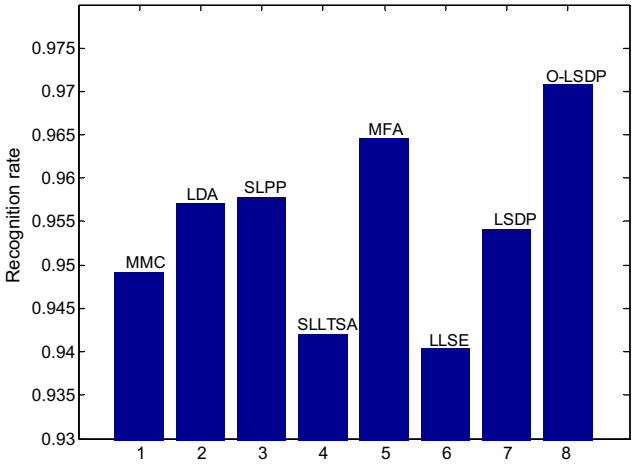

Fig. 7. Performance comparison of best mean recognition rates using MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP on the ORL database.

Table 2
The maximal average recognition rates and the corresponding standard deviations (%) with the reduced dimensions for BASELINE, MMC, LDA, SLPP, SLLTSA, MFA, LLSE, LSDP, and O-LSDP on the ORL database.

Method	3 Train	4 Train	5 Train	6 Train	7 Train
Baseline	77.14 ± 2.23	82.48 ± 2.22	86.90 ± 2.12	89.34 ± 2.55	90.92 ± 2.26
MMC	83.45 ± 1.90 (39)	88.83 ± 2.15 (38)	92.47 ± 2.13 (37)	93.88 ± 2.34 (34)	94.92 ± 1.77 (36)
LDA	85.86 ± 1.68 (38)	90.33 ± 1.66 (39)	93.23 ± 1.91 (39)	94.62 ± 1.96 (39)	95.71 ± 1.86 (39)
SLPP	85.84 ± 1.62 (39)	90.25 ± 1.63 (39)	93.28 ± 1.92 (39)	94.66 ± 1.91 (39)	95.79 ± 1.90 (39)
SLLTSA	81.00 ± 2.13 (81)	86.02 ± 2.06 (83)	90.30 ± 1.64 (95)	92.34 ± 2.6 (99)	94.21 ± 1.92 (111)
MFA	81.38 ± 2.79 (34,2,305)	89.85 ± 2.26 (42,3,319)	93.47 ± 2.35 (43,4,215)	95.44 ± 1.69 (44,5,246)	96.46 ± 1.97 (51,6,242)
LLSE	77.54 ± 3.25 (32)	84.92 ± 2.93 (32)	89.43 ± 2.22 (33)	92.44 ± 2.23 (27)	94.04 ± 2.50 (30)
LSDP	85.29 ± 1.55 (39)	89.88 ± 1.70 (39)	92.80 ± 1.79 (38)	94.50 ± 1.94 (39)	95.42 ± 2.00 (39)
O-LSDP	**86.73 ± 1.55 (40)**	**93.15 ± 1.72 (43)**	**95.40 ± 1.37 (37)**	**96.22 ± 1.51 (35)**	**97.08 ± 2.12 (34)**

For MFA, the first numbers in the parentheses are the selected subspace dimensions, the second and the third numbers are the parameters k_1 and k_2, respectively.

Fig. 8. Sample images from one person in the extended Yale B database.

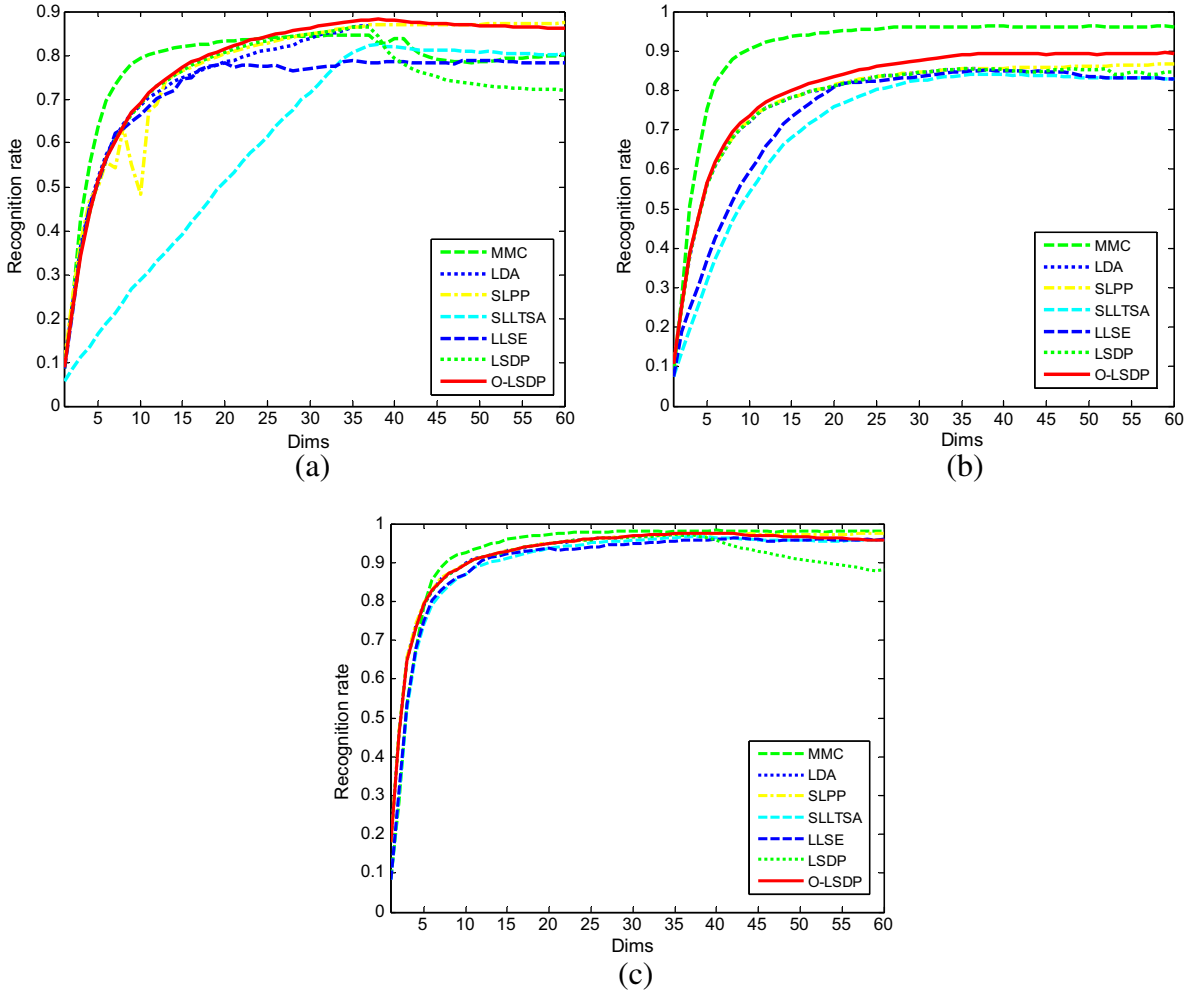

Fig. 9. The recognition rates of MMC, LDA, SLPP, SLLTSA, LLSE, LSDP, and O-LSDP versus the dimensions on the extended Yale B database. (a) 10 samples for training. (b) 30 samples for training. (c) 50 samples for training.

Table 3
The maximal average recognition rates and the corresponding standard deviations (%) with the reduced dimensions for BASELINE, MMC, LDA, SLPP, SLLTSA, LLSE, LSDP, and O-LSDP on the extended Yale B database.

Method	10 Train	30 Train	50 Train
Baseline	53.44 ± 0.82	77.39 ± 0.98	84.22 ± 1.46
MMC	84.70 ± 1.30 (37)	96.38 ± 0.47 (40)	98.27 ± 0.41 (40)
LDA	86.84 ± 1.13 (37)	85.46 ± 1.26 (37)	97.28 ± 0.56 (35)
SLPP	87.38 ± 1.06 (60)	86.70 ± 1.11 (39)	97.49 ± 0.56 (60)
SLLTSA	85.25 ± 1.57 (103)	84.09 ± 1.47 (38)	96.69 ± 0.68 (38)
LLSE	85.72 ± 1.34 (68)	85.19 ± 1.42 (38)	96.35 ± 0.52 (43)
LSDP	86.54 ± 1.14 (37)	85.44 ± 1.21 (36)	97.19 ± 0.53 (37)
O-LSDP	**89.89 ± 0.99 (231)**	**97.24 ± 0.43 (957)**	**98.55 ± 0.43 (775)**

manifold structure (He et al., 2005a,b). In addition, compared to other manifold learning-based methods, O-LSDP encodes more discriminant information in the reduced feature subspace by more faithfully preserving local geometry and incorporating the class information.

2. O-LSDP significantly outperforms SLLTSA, irrespective of the variations in training sample size. There are two reasons contributing to this phenomenon. On the one hand, although both SLLTSA and O-LSDP use local tangent space coordinates as their local geometry, they are intrinsically different in the global alignment stage. In SLLTSA, an affine transformation is used to align the local coordinates while in O-LSDP, smooth spline functions are constructed to perform the whole alignment. Compared to the affine transformation, splines can more faithfully preserve local geometry, which is directly related to the discriminant power (He et al., 2005a,b). On the other hand, orthogonalization contributes to the noise removal (Ye, 2005) and more locality preserving power.
3. MFA has comparative recognition rates with O-LSDP when training sample size is 6 or 7 in ORL database. This is because MFA can also effectively capture both the local geometry and the discriminant information of data by setting k_1 and k_2 suitably. However, it is necessary to traverse all possible values of k_1 and k_2 for model selection. Therefore, the computational cost of the MFA algorithm grows quickly as the sample size is increased.

5. Conclusions

In this paper, we have introduced a novel linear dimensionality reduction algorithm for face recognition, called orthogonal linear local spline discriminant projection (O-LSDP). The most prominent property for O-LSDP is to successfully manage the trade-off between the discriminant power and local geometrical structure hidden in the data. We have applied our algorithm to face recognition. The experimental results on Yale, ORL, and Extended Yale B databases show that the proposed method is indeed effective and efficient.

Acknowledgments

This work is supported by the grants of the National Science Foundation of China, Nos. 60705007, 60805021, 60975005, 60873012, 60905023 and 60872113, the grant from the National Basic Research Program of China (973 Program, No. 2007CB311002), the grants from the National High Technology Research and Development Program of China (863 Program, No. 2007AA01Z167), and the Knowledge Innovation Program of the Chinese Academy of Sciences. The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

References

Belhumeour, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. PAMI 19 (7), 711–720.
Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15 (6), 1373–1396.
Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M., 2003. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In: Proc. NIPS, vol. 16, pp. 177–184.
Brand, M., 2003. Charting a manifold. In: Proc. NIPS, vol. 15, pp. 961–968.
Cai, D., He, X., Han, J., 2005. Using graph model for face analysis. Technical Report No. 2636, Dept. of Computer Science, Univ. of Illinois at Urbana-Champaign.
Chin, T.-J., Suter, D., 2008. Out-of-sample extrapolation of learned manifolds. IEEE Trans. PAMI 30 (9), 1547–1556.
Coifman, R.R., Lafon, S., 2006. Diffusion maps. Appl. Comput. Harmonic Anal. 21, 5–30.
DeCoste, D., 2001. Visualizing Mercer kernel feature spaces via kernelized locally linear embeddings. In: Proc. ICNIP, pp.14–18.
Donoho, D.L., 2000. High-dimensional data analysis: The curses and blessings of dimensionality. In: Proc. AMS Math. Challenges of the 21st Century.
Donoho, D., Grimes, C., 2003. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. 100 (10), 5591–5596.
Duchene, J., Leclercq, S., 1988. An optimal transformation for discriminant and principal component analysis. IEEE Trans. PAMI 10 (6), 978–983.
Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley & Sons, New York.
Geng, X., Zhan, D.C., Zhou, Z.H., 2005a. Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Trans. Systems Man Cybern. 35 (6), 1098–1107.
He, X., Niyogi, P., 2003. Locality preserving projections. In: Proc. NIPS.
He, X., Cai, D., Yan, S., Zhang, H., 2005a. Neighborhood preserving embedding. In: Proc. ICCV, pp. 1208–1213.
He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J., 2005b. Face recognition using Laplacian faces. IEEE Trans. PAMI 27 (3), 328–340.
Huang, D.S., 1996. Systematic Theory of Neural Networks for Pattern Recognition. Publishing House of Electronic Industry of China, Beijing.
Huang, D.S., 1998. The local minima free condition of feedforward neural networks for outer-supervised learning. IEEE Trans. Systems Man Cybern. 28B (3), 477–480.
Huang, D.S., 1999. Radial basis probabilistic neural networks: Model and application. Internat. J. Pattern Recognition Artif. Intell. 13 (7), 1083–1101.
Huang, D.S., Ma, S.D., 1999. Linear and nonlinear feedforward neural network classifiers: A comprehensive understanding. J. Intell. Systems 9 (1), 1–38.
Jolliffe, I.T., 1989. Principal Component Analysis. Springer, New York.
Kokiopoulou, E., Saad, Y., 2007. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. IEEE Trans. PAMI 29 (12), 2143–2156.
Lafon, S., Lee, A.B., 2006. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning and data set parameterization. IEEE Trans. PAMI 28 (9), 1393–1403.
Law, M.H., Jain, A.K., 2006. Incremental nonlinear dimensionality reduction by manifold learning. IEEE Trans. PAMI 28 (3), 377–391.
Li, H., Jiang, T., Zhang, K., 2006. Efficient and robust feature extraction by maximum margin criterion. IEEE Trans. Neural Networks 17 (1), 157–165.
Lin, T., Zha, H., 2008. Riemannian manifold learning. IEEE Trans. PAMI 30 (5), 796–809.
Lin, T., Zha, H., Lee, S., 2006. Riemannian manifold learning for nonlinear dimensionality reduction. In: Proc. ECCV, pp. 44–55.
Pan, Y., Ge, S.S., Mamun, A.A., 2008. Weighted locally linear embedding for dimension reduction. Pattern Recognition 42 (5), 798–811.
Pang, Y., Liu, Z., Yu, N., 2006. A new nonlinear feature extraction method for face recognition. Neurocomputing 69, 949–953.
Ridder, D., Kouropteva, O., Okun, O., Pietikäinen, M., Duin, R.P.W., 2003. Supervised locally linear embedding. In: Proc. ICANN/ICONIP 2714, pp. 333–341.
Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.
Saul, L.K., Roweis, S.T., 2003. Think globally, fit locally: Unsupervised learning of low dimensional manifold. J. Machine Learn. Res. 4, 119–155.
Silva, V., Tenenbaum, J., 2003. Global versus local methods in nonlinear dimensionality reduction. In: Proc. NIPS, vol. 15, pp. 705–712.
Tenenbaum, J., de Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.
Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cognitive Neurosci. 3 (1), 71–86.
Weinberger, K., Saul, L., 2004. Unsupervised learning of image manifolds by semidefinite programming. In: Proc. CVPR, vol. 2, pp. 988–995.
Xiang, S.M., Nie, F.P., Zhang, C.S., 2009. Nonlinear dimensionality reduction with local spline embedding. IEEE Trans. Knowledge Data Eng. 21 (9), 1285–1298.
Xiang, S.M., Nie, F.P., Zhang, C.S., Zhang, C.X., 2006. Spline embedding for nonlinear dimensionality reduction. In: Proc. ECML, pp. 825–832.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., Lin, S., 2007. Graph embedding and extension: A general framework for dimensionality reduction. IEEE Trans. PAMI 29 (1), 40–51.

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

Zhang, J., Shen, H., Zhou, Z.-H., 2004. Unified locally linear embedding and linear discriminant analysis algorithm for face recognition. Lecture Notes in Computer Science. Springer, Berlin.

Zhang, T.H., Yang, J., Zhao, D.L., Ge, X.L., 2007. Linear local tangent space alignment and application to face recognition. Neurocomputing 70, 1547–1553.

Zhang, Z., Zha, H., 2005. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. SIAM J. Sci. Comput. 26 (1), 313–338.

Zhang, Z., Zhao, L., 2007. Probability-based locally linear embedding for classification. In: Proc. FSKD, pp. 243–247.