Neural Process Lett (2011) 33:235–250
DOI 10.1007/s11063-011-9175-z

Orthogonal Complete Discriminant Locality Preserving Projections for Face Recognition

Gui-Fu Lu · Zhong Lin · Zhong Jin

Published online: 20 April 2011

Abstract In this paper, we propose a novel orthogonal complete discriminant locality preserving projections for facial feature extraction and recognition (OCDLPP). All training samples are projected into the range of a so-called *locality preserving total scatter* to reduce dimensionality without loss of discriminative information. The transformation matrix of OCDLPP is orthogonal and is found simultaneously using QR decomposition technique. Moreover, a feasible and effective procedure is proposed to alleviate the computational burden of high dimensional matrix for typical face image data. Experiments results on the ORL, Yale, FERET and PIE face databases show the effectiveness of the proposed OCDLPP.

Keywords Face recognition · Locality preserving projections · Orthogonal complete discriminant locality preserving · Small size sample problems · Feature extraction

1 Introduction

Face recognition has attracted many researchers in the area of pattern recognition, machine learning, and computer vision because of its immense application potential. Numerous methods have been proposed in the last two decades [1,2]. One of the most successful and well-studied techniques to face recognition is the appearance-based method. In an appearance-based technique, a face image with size $m \times n$ is perceived as a point in a $m \times n$-dimensional image space. In practice, however, these $m \times n$-dimensional spaces are too large to allow robust and fast recognition. Dimensionality reduction is an effective approach to deal with this problem. The most well-know dimensionality reduction methods are principal component analysis (PCA) [3] and linear discriminant analysis (LDA) [4].

G.-F. Lu · Z. Lin · Z. Jin
School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, 210094 Jiangsu, China

G.-F. Lu (✉)
School of Computer Science and Information, Anhui Polytechnic University, Wuhu, 241000 Anhui, China
e-mail: luguifu_jsj@163.com

PCA aims at finding a set of mutually orthogonal basis vectors that captures the directions of maximum variance in the data. PCA exhibits optimality in the sense of reconstruction error. However, PCA is not ideal for classification since it does not make use of any output class information in computing the principle components. LDA is a supervised learning algorithm which derives a projection basis that separates different class as far as possible and compresses the same class as compact as possible. Unfortunately, it cannot be applied directly to the small size sample (SSS) problem, e.g. face recognition, because the within-class scatter matrix is singular. Numerous methods have been proposed to solve this problem in the last decades, including PCA + LDA [4], Regularized LDA (RLDA) [5], null space based LDA (NLDA) [6,7], orthogonal LDA (OLDA) [8], QR/GSVD-based LDA (LDA/QR-GSVD) [9–12]. In addition, Tao et al. [13] proposed a geometric mean based subspace selection algorithm, which maximizing the geometric mean of Kullback–Leibler (KL) divergences between different class pairs. This geometric mean method can improve the discriminant performance of LDA. In [14], Bian and Tao applied the harmonic mean to replace the geometric mean for subspace selection and empirically demonstrated the effective of harmonic mean in dealing with the class separation problem.

Recent studies have shown the face images possibly reside on a nonlinear submanifold [15–17]. Many manifold-based learning algorithms have been proposed, e.g. Isomap [15], locally linear embedding (LLE) [16], and Laplacian Eigenmap [17]. Recently, Yan et al. [18] introduced a general framework for dimensionality reduction, called graph embedding, where a large number of popular dimensionality reduction algorithms can be considered as special cases within the framework. Similarly, a patch alignment framework [19] devotes to show the existing algorithms are different in the patch reconstruction while sharing an identical whole alignment stage. In [20,21], Zhou et al. presented another unified framework, called manifold elastic net (MEN), for sparse dimensionality reduction. The algorithm can obtain a sparse solution to supervised subspace learning by using $L1$ manifold regularization. Especially in the cases of small training sets and lower-dimensional subspaces, it achieves better classification performances against traditional subspace learning algorithms. Geng et al. [22] presented an ensemble manifold regularization method to approximate the intrinsic manifold structure by combining several initial guesses. Manifold regularization was also used in sliced inverse regression [23] to get better classification performances and visualization results. In [24], Si et al. proposed a Bregman divergence-based regularization framework, which minimizes the Bregman divergences between the distribution of training samples and that of testing samples in the selected subspace and the performances of traditional subspace learning methods are boosted when training and testing samples are not independent and identically distributed.

Recently, LPP [25,26] has been given more attention in pattern recognition. Different from PCA and LDA which effectively see only the Euclidean structure of face space, LPP finds an embedding that preserving local information, and obtains a face subspace that best detects the essential face manifold structure. Based on LPP, the Laplacianfaces was further developed for face recognition [26], giving encouraging performance. One limitation of LPP is that it is still an unsupervised method. To consider the discriminant information of recognition task, several locality preserving discriminant analysis methods have been mentioned in recent years. Yu et al. [27] presented a discriminant locality preserving projections (DLPP) method to improve the classification performance of LPP. Similar to LDA, DLPP also suffers from the SSS problem. So PCA approach, which discards some useful discriminatory information, is often used before DLPP. Yang et al. [28] proposed a null space discriminant locality preserving projections (NDLPP) algorithms. However, NDLPP merely utilizes the discriminant information in the null space of the *locality preserving within-class scatter*.

As the orthogonal projection is of desirable property and often demonstrates good performance empirically [8,29–36], many orthogonal methods, e.g. orthogonal LDA (OLDA) [8], orthogonal neighborhood preserving discriminant analysis (ONPDA) [29], orthogonal LPP [36] and trace ratio based orthogonal discriminant analysis (TRODA) [30], have been proposed for feature extraction.

In this paper, in order to overcome the SSS problem encountered by DLPP and produce orthogonal discriminant vectors, we propose a novel orthogonal complete discriminant locality preserving projections (OCDLPP) method. OCDLPP not only comes into the character of DLPP that encodes both the geometrical and discriminant structure of the data manifold, but also produces orthogonal discriminant vectors, i.e. the transformation matrix of OCDLPP is orthogonal. The set of orthogonal discriminant vectors of OCDLPP is computed simultaneously via QR decomposition technique. Moreover, an efficient algorithm for implementing OCDLPP is also developed without any loss of effective discriminatory information and the SSS problem encountered by DLPP is overcome naturally. Extensive experimental studies on the ORL, Yale, FERET and PIE face databases show the effectiveness of the proposed OCDLPP method.

The organization of the rest of this paper is as follows. In Sect. 2, we review briefly the discriminant locality preserving projections (DLPP). In Sect. 3, we propose the idea and describe the new method in detail. Some related works are also compared with OCDLPP in this section. In Sect. 4, experiments with face images data are presents to demonstrate the effectiveness of the OCDLPP algorithms. Conclusions are made in Sect. 5.

2 Outline of Discriminant Locality Preserving Projections

A set of face image sample $\{\mathbf{x}_i\}$ can be represented as a $M \times N$ matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, where M is the number of pixels in the images and N is the number of samples. Each face image $\mathbf{x}_i$ belongs to one of the C face classes $\{X_1, \ldots, X_C\}$. DLPP tries to maximize an objective function as follows:

$$\frac{\sum_{i,j=1}^{C} (\mathbf{m}_i - \mathbf{m}_j) B_{ij} (\mathbf{m}_i - \mathbf{m}_j)^T}{\sum_{c=1}^{C} \sum_{i,j=1}^{n_c} (\mathbf{y}_i^c - \mathbf{y}_j^c) W_{ij}^c (\mathbf{y}_i^c - \mathbf{y}_j^c)^T} \tag{1}$$

where n_c is the number of samples in the cth class, $\mathbf{y}_i^c$ represents the ith projected vector in the cth class, $\mathbf{m}_i$ and $\mathbf{m}_j$ is separately the mean projected vector for the ith class and jth class, i.e. $\mathbf{m}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{y}_k^i$ and $\mathbf{m}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} \mathbf{y}_k^j$, where n_i and n_j is the number of samples in the ith and jth class, separately. W_{ij}^c represents the elements of within-class weight matrix and $W_{ij}^c = \exp(-\|\mathbf{x}_i^c - \mathbf{x}_j^c\|^2/\sigma^2)$, and B_{ij} represents the elements of between-class weight matrix and $B_{ij} = \exp(-\|\mathbf{f}_i - \mathbf{f}_j\|^2/\sigma^2)$, where σ is an empirically determined parameter, $\mathbf{x}_i^c$ represents the ith vector in the cth class, $\mathbf{f}_i$ is the mean of the ith class, i.e. $\mathbf{f}_i = (1/n_i) \sum_{k=1}^{n_i} \mathbf{x}_k^i$. Thus, the between-class weight matrix is $B = [B_{ij}](i, j = 1, 2, \ldots, C)$, the within-class weight matrix is $W = diag\{W^{(1)}, \ldots, W^{(C)}\}$, where $W^{(i)} = [W_{jk}^{(i)}](j, k = 1, 2, \ldots, n_i)$. It is clear that both B and W are symmetric positive semi-definite matrices.

Suppose that the mapping from $\mathbf{x}_i$ to $\mathbf{y}_i$ is A, i.e. $\mathbf{y}_i = A^T \mathbf{x}_i$, then, the objective function (1) can be rewritten as

$$J_1(A) = \frac{A^T F H F^T A}{A^T X L X^T A} \tag{2}$$

where L and H are Laplacian matrices. $L = D - W$, $D = diag\{D_1, \ldots, D_C\}$, D_i is a diagonal matrix and its elements are column (or row) sum of $W^{(i)}$; $H = E - B$, E is a diagonal matrix and its elements are column (or row) sum of B. $F = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_C]$.

Now we would give the following definitions:
locality preserving within-class scatter: $S_w^L = XLX^T$;
locality preserving between-class scatter: $S_b^L = FHF^T$;
locality preserving total scatter: $S_t^L = S_b^L + S_w^L$.

It is clear that S_w^L, S_b^L and S_t^L are all symmetric positive semi-definite matrices. The transformation matrix $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k]$ that maximizes the objective function (2) can be obtained by solving the generalized eigenvalues problem:

$$(FHF^T)\mathbf{a}_i = \lambda_i (XLX^T)\mathbf{a}_i, \lambda_1 \geq \lambda_2 \geq \cdots \lambda_k \tag{3}$$

or

$$S_b^L \mathbf{a}_i = \lambda_i S_w^L \mathbf{a}_i, \lambda_1 \geq \lambda_2 \geq \cdots \lambda_k \tag{4}$$

DLPP requires that matrix XLX^T be nonsingular. For many applications involving the SSS problem this matrix is singular. In fact, as long as the dimension of sample M is greater than the number of samples N, XLX^T must be singular. This can be induced from simple matrix computation knowledge, $rank(XLX^T) \leq rank(L) \leq N < M$. Thus DLPP can't be applied directly. So PCA approach, which discards some useful discriminatory information, is used before DLPP.

3 Orthogonal Complete Discriminant Locality Preserving Projections

3.1 Theoretical Derivation of OCDLPP

It is obvious the objective function (2) is equivalent to

$$J_2(A) = \frac{A^T S_b^L A}{A^T S_t^L A} \tag{5}$$

The OCDLPP algorithm is to solve the following optimization problem:

$$A = \arg \max_{A^T A = I} tr\left(\left(A^T S_t^L A\right)^+ A^T S_b^L A\right) \tag{6}$$

where I denotes the identity matrix, $()^+$ denotes the Moore and Penrose generalized inverse of a matrix and $tr(\cdot)$ denotes the trace of a matrix.

Firstly, we discuss how to solve the following optimization problem:

$$G = \arg \max tr\left(\left(G^T S_t^L G\right)^+ G^T S_b^L G\right) \tag{7}$$

The matrices S_w^L and S_b^L are both positive semi-definite, so the intersection of their null spaces is equal to the null space of S_t^L, namely, $\{\mathbf{x} | S_t^L \mathbf{x} = 0\}$. As the null space of S_t^L does not contain discriminating information for the training data ($\mathbf{x}^T S_w^L \mathbf{x} = 0$ and $\mathbf{x}^T S_b^L \mathbf{x} = 0$), it may be removed from the solution space without accuracy. Assume that the eigenvalue decomposition of matrix S_t^L is

$$S_t^L = U \Lambda U^T \tag{8}$$

where $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_{m'}\}$, $\lambda_i > 0, i = 1, 2, \ldots, m'$, and m' is the number of positive singular values of S_t^L, $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m'}]$ is the eigenvectors of S_t^L corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{m'}$. Then the solution is constrained to lie within the space spanned by the column vector of U, namely, $G = UP$, $P \in R^{m' \times k}$. Therefore, Eq. 5 can be changed to

$$J_2(P) = \frac{P^T \tilde{S}_b^L P}{P^T \tilde{S}_t^L P} = \frac{P^T \tilde{S}_b^L P}{P^T \Lambda P} \tag{9}$$

where $\tilde{S}_b^L = U^T S_b^L U$ and $\tilde{S}_t^L = U^T S_t^L U = \Lambda$. Then, the denominator of the objective function (9) is always positive for non-zero P, i.e. $\tilde{S}_t^L$ is positive definite. The transformation matrix $P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k]$ that maximizes the objective function (9) can be obtained by solving the eigenvalues problem:

$$\Lambda^{-1} \tilde{S}_b^L \mathbf{p}_i = \lambda_i \mathbf{p}_i, \lambda_1 \geq \lambda_2 \geq \cdots \lambda_k \tag{10}$$

Then the optimal projection matrix is given by $G = UP$, where $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m'}]$ and $P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k]$

Theorem 1 *The columns of $G = UP$ are eigenvectors of $(S_t^L)^+ S_b^L$.*

Proof Let $\mathbf{p} \in R^{m' \times 1}$ to be the eigenvector of (10) corresponding to the eigenvalue λ, we have

$$\Lambda^{-1} \tilde{S}_b^L \mathbf{p} = \lambda \mathbf{p} \tag{11}$$

Because of $\tilde{S}_b^L = U^T S_b^L U$, (11) can be changed to

$$U \Lambda^{-1} U^T S_b^L (U\mathbf{p}) = \lambda (U\mathbf{p}) \tag{12}$$

Because of $(S_t^L)^+ S_b^L = U \Lambda^{-1} U^T S_b^L$, $U\mathbf{p}$ is an eigenvectors of $(S_t^L)^+ S_b^L$. □

According to Theorem 1, we can conclude that $G = UP$ is the optimal discriminant transformation of (7). Now we discuss how to obtain the optimal discriminant transformation of (6). Let $V \in R^{k \times k}$ is an arbitrary nonsingular matrix and $\tilde{G} = GV$, then we have

Theorem 2 $tr\left((G^T S_t^L G)^+ G^T S_b^L G\right) = tr((\tilde{G}^T S_t^L \tilde{G})^+ \tilde{G}^T S_b^L \tilde{G})$

Proof

$$\begin{aligned}
& tr\left(\left(\tilde{G}^T S_t^L \tilde{G}\right)^+ \tilde{G}^T S_b^L \tilde{G}\right) \\
&= tr\left(\left((GV)^T S_t^L (GV)\right)^+ (GV)^T S_b^L (GV)\right) \\
&= tr\left(\left(V^T G^T S_t^L GV\right)^+ V^T G^T S_b^L GV\right)
\end{aligned}$$

Since $V^T G^T S_t^L G V = V^T P^T \Lambda P V$ and V are nonsingular, we have

$$
\begin{aligned}
& tr\left(\left(V^T G^T S_t^L G V\right)^+ V^T G^T S_b^L G V\right) \\
&= tr\left(V^- \left(G^T S_t^L G\right)^+ \left(V^T\right)^- V^T G^T S_b^L G V\right) \\
&= tr\left(\left(G^T S_t^L G\right)^+ G^T S_b^L G\right)
\end{aligned}
$$

□

According to Theorem 2, $\tilde{G} = GV$ is the optimal discriminant transformation of (7) for nonsingular matrix V. From matrix computation knowledge, we know that G could be QR-decomposed as $G = \tilde{Q}\tilde{R}$, where $\tilde{Q} \in R^{M\times k}$ has the orthonormal columns and $\tilde{R} \in R^{k\times k}$ is an upper triangular matrix. Now we choose $V = \tilde{R}^{-1}$ so that the columns of $GV = G\tilde{R}^{-1} = \tilde{Q}$ are orthogonal to each other. Let $A = \tilde{Q}$, then A is the optimal discriminant transformation of (6).

3.2 Computational Considerations of OCDLPP

However, S_t^L is an $M \times M$ matrix where M is very large as typically the case when dealing with face image data. In this case, it will be very time-consuming to compute the eigenvalue decomposition of S_t^L as the computational complexity is $O(M^3)$. To reduce the computational demand, in this subsection, we present an efficient and stable algorithm for performing OCDLPP.

The Laplacian matrices L and H are always real symmetric positive semi-definite, so L and H can be decomposed as follows:

$$L = V_L \Lambda_L V_L^T,\ H = V_H \Lambda_H V_H^T \tag{13}$$

where Λ_L is the eigenvalue matrix of L, i.e. $\Lambda_L = diag\{\lambda_1^L, \lambda_2^L, \ldots, \lambda_N^L\}$, the column of V_L are the orthogonal eigenvectors of corresponding to eigenvalues of L; Λ_H is the eigenvalue matrix of H, i.e. $\Lambda_H = diag\{\lambda_1^H, \lambda_2^H, \ldots, \lambda_C^H\}$, the column of V_H are the orthogonal eigenvectors of corresponding to eigenvalues of H.

It is easy to know that all the eigenvalues of both L and H are non-negative since L and H are both real symmetric semi-positive definite matrices. Consequently, S_w^L, S_b^L and S_t^L can be rewritten as

$$S_w^L = XLX^T = H_w H_w^T \tag{14}$$
$$S_b^L = FHF^T = H_b H_b^T \tag{15}$$
$$S_t^L = S_w^L + S_b^L = H_t H_t^T \tag{16}$$

where $H_w = XV_L\Lambda_L^{1/2} \in R^{M\times N}$, $H_b = FV_H\Lambda_H^{1/2} \in R^{M\times C}$ and $H_t = [H_w H_b] \in R^{M\times(N+C)}$. Assume that the thin singular value decomposition of matrix H_t is

$$H_t = U\Lambda_t Q^T \tag{17}$$

where Λ_t is the singular value matrix of H_t and $\Lambda = \Lambda_t^2$, $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m'}]$ is the left singular vector matrix of H_t, and Q is the right singular vector matrix of H_t. So we can obtain the eigenvectors $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m'}]$ of S_t^L by the thin singular value decomposition of matrix H_t. Note that the size of H_w, H_b and H_t are much smaller than that of S_t^L, since usually $N \ll M$ and $C \ll M$.

Now, the algorithmic procedure of OCDLPP is formally summarized as follows.

(1) Construct the within-class weight matrix W and between-class weight matrix B. Then calculate the within-class Laplacian matrix L and between-class Laplacian matrix H.
(2) Compute H_w, H_b and H_t by (13) (14) (15) and (16).
(3) Perform the thin singular value decomposition of matrix H_t as (17).
(4) Compute $\tilde{S}_b^L = U^T S_b^L U$ and $\tilde{S}_t^L = U^T S_t^L U = \Lambda$.
(5) Solve the eigenvalue problem $\Lambda^{-1}\tilde{S}_b^L \mathbf{p} = \lambda \mathbf{p}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ be the k largest eigenvalues of $\Lambda^{-1}\tilde{S}_b^L$ and $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k$ be the associated eigenvectors.
(6) Let $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m'}]$ and $P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k]$, then decompose $G = UP$ using QR decomposition technique as $G = \tilde{Q}\tilde{R}$, where $\tilde{Q} \in R^{M \times k}$ has the orthonormal columns and $\tilde{R} \in R^{k \times k}$ is an upper triangular matrix.
(7) Let $A = \tilde{Q}$ and A is the optimal discriminant transformation of (6).

It is worthwhile to highlight several valuable properties of the proposed OCDLPP here:

(1) The dimension of feature space is first greatly reduced without loss on discriminative information by removing the null space of S_t^L and the SSS problem is overcome naturally. However, LDA, LPP, OLPP and DLPP involve a preceding PCA stage to avoid the SSS problem and some useful discriminatory information is discarded.
(2) The computational complexities for eigenvalue decomposition of L and H are $O(N^3)$ and $O(C^3)$ respectively and the computational complexity for singular value decomposition of H_t is $O(M(N+C)^2)$. Thus, the computational savings are great in face recognition since the computational complexity for eigenvalue decomposition of S_t^L is $O(M^3)$ and usually $N \ll M$ and $C \ll M$.
(3) The optimal discriminant transformation of OCDLPP is orthogonal and is obtained simultaneously using QR decomposition technique.

3.3 Comparison with Related Works

Recently, several orthogonal methods, e.g. orthogonal LDA (OLDA) [8], orthogonal LPP (OLPP) [36] and trace ratio based orthogonal discriminant analysis (TRODA) [30], have been proposed for feature extraction. In this section, we discuss the relationships among OLDA, OLPP, TRODA and our proposed OCDLPP.

The trace ratio based orthogonal discriminant analysis (TRODA) [30] is to solve the optimization problem as follows:

$$A = \arg \max_{A^T A = I} \frac{tr(A^T S_p A)}{tr(A^T S_l A)} \tag{18}$$

where S_p and S_l are both positive semi-definite matrices. However, TRODA lacks a direct and globally optimal solution. In [30], Wang et al. proposed an iterative procedure to solve this problem. However, the solution proposed by Wang et al. may be locally optimal and it needs much computational cost. The objective function of our proposed method is a ratio trace problem which has a close-form solution. Then OCDLPP does not have the convergence problem and local optimum problem.

The OLDA [8] algorithm, which is also a ratio trace problem, is to solve optimization problem as follows:

$$A = \arg \max_{A^T A = I} tr\left(\left(A^T S_t A\right)^{-1} A^T S_b A\right) \tag{19}$$

where S_t is the total scatter matrix and S_b is the between-class scatter matrix. Similar to OCDLPP, Ye [8] also uses QR decomposition technique to compute the orthogonal transformation matrix of OLDA simultaneously. However, their optimization problems are different. OLDA is a global technique that perceives only the Euclidean structure and can not discover the intrinsic manifold from the samples. On the other hand, OCDLPP considers the local neighboring geometry that corresponds to the nonlinear data structure and uses sample similarity weight to capture the embedding sample manifold structure.

Cai et al. [36] proposed an OLPP algorithm. They used a step-by-step procedure to obtain a set of orthogonal projection $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\}$. After calculating the first $k-1$ projections $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1}\}$, the kth projection $\mathbf{a}_k$ is calculated by solving the following optimization problem:

$$\mathbf{a}_k = \arg \min_{A_{k-1}^T \mathbf{a}_k = 0} \frac{\mathbf{a}_k^T X L X^T \mathbf{a}_k}{\mathbf{a}_k^T X D X^T \mathbf{a}_k} \tag{20}$$

where D is the degree matrix on the graph and L is a Laplacian matrix. Define

$$A_{k-1} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1}] \tag{21}$$

$$B_{k-1} = A_{k-1}^T (X D X^T)^{-1} A_{k-1} \tag{22}$$

Then $\mathbf{a}_k$ is the eigenvector of

$$M_k = \left\{I - (X D X^T)^{-1} A_{k-1} (B_{k-1})^{-1} A_{k-1}^T\right\} \cdot (X D X^T)^{-1} X L X^T \tag{23}$$

associated with the smallest eigenvalue. The step-by-step procedure makes the algorithm computationally more expensive and makes the objective with regard to A to be optimized not clear.

Suppose the trace ratio based orthogonal LPP is to solve the following optimization problem:

$$A = \arg \min_{A_{k-1}^T A_k = I} \frac{tr(A^T X L X^T A)}{t^r(A^T X D X^T A)} \tag{24}$$

Then if we use a step-by-step procedure to calculate A, we have the following theorem.

Theorem 3 *Suppose*

$$A_{k-1} = \arg \min_{A_{k-1}^T A_{k-1} = I} \frac{tr\left(A_{k-1}^T X L X^T A_{k-1}\right)}{tr\left(A_{k-1}^T X D X^T A_{k-1}\right)} \tag{25}$$

and the orthogonal transformation matrix $A_{k-1} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1}]$ *has been calculated. Let the kth projection* $\mathbf{a}_k$ *is calculated by solving the following optimization problem:*

$$\mathbf{a}_k = \arg \min_{A_{k-1}^T \mathbf{a}_k = 0} \frac{tr(A^T X L X^T A)}{tr(A^T X D X^T A)} \tag{26}$$

Then $\mathbf{a}_k$ *is also the eigenvector of*

$$M_k = \left\{I - (X D X^T)^{-1} A_{k-1} (B_{k-1})^{-1} A_{k-1}^T\right\} \cdot (X D X^T)^{-1} X L X^T \tag{27}$$

associated with the smallest eigenvalue.

Proof Equation 26 can be solved by Lagrangian multiplier method:

$$C^{(k)} = \frac{tr(A^T XLX^T A)}{tr(A^T XDX^T A)} - \sum_{i=1}^{k-1} \mu_i \mathbf{a}_k^T \mathbf{a}_i = \frac{\sum_{i=1}^{k-1} \mathbf{a}_i^T XLX^T \mathbf{a}_i + \mathbf{a}_k^T XLX^T \mathbf{a}_k}{\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k} - \sum_{i=1}^{k-1} \mu_i \mathbf{a}_k^T \mathbf{a}_i \tag{28}$$

We set the partial derivative of $C^{(k)}$ to $\mathbf{a}_k$ to zero and obtain

$$2\frac{XLX^T \mathbf{a}_k \left(\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k\right) - XDX^T \mathbf{a}_k \left(\sum_{i=1}^{k-1} \mathbf{a}_i^T XLX^T \mathbf{a}_i + \mathbf{a}_k^T XLX^T \mathbf{a}_k\right)}{\left(\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k\right)^2} - \sum_{i=1}^{k-1} \mu_i \mathbf{a}_i = 0 \tag{29}$$

Let

$$\lambda = \frac{tr(A^T XLX^T A)}{tr(A^T XDX^T A)} \tag{30}$$

Multiplying the left side of (29) by $\mathbf{a}_k^T$, we obtain

$$\frac{\mathbf{a}_k^T XLX^T \mathbf{a}_k}{\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k} - \lambda \frac{\mathbf{a}_k^T XDX^T \mathbf{a}_k}{\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k} = 0 \tag{31}$$

and

$$\lambda = \frac{\mathbf{a}_k^T XLX^T \mathbf{a}_k}{\mathbf{a}_k^T XDX^T \mathbf{a}_k} \tag{32}$$

Multiplying the left side of (29) successively by $\mathbf{a}_1^T (XDX^T)^{-1}, \ldots, \mathbf{a}_{k-1}^T (XDX^T)^{-1}$, we now obtain a set of $k-1$ equations:

$$\begin{aligned}
&2\frac{\mathbf{a}_1^T (XDX^T)^{-1} XLX^T \mathbf{a}_k}{\left(\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k\right)} = \sum_{i=1}^{k-1} \mu_i \mathbf{a}_1^T (XDX^T)^{-1} \mathbf{a}_i \\
&2\frac{\mathbf{a}_2^T (XDX^T)^{-1} XLX^T \mathbf{a}_k}{\left(\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k\right)} = \sum_{i=1}^{k-1} \mu_i \mathbf{a}_2^T (XDX^T)^{-1} \mathbf{a}_i \\
&\cdots\cdots \\
&2\frac{\mathbf{a}_{k-1}^T (XDX^T)^{-1} XLX^T \mathbf{a}_k}{\left(\sum_{i=1}^{k-1} \mathbf{a}_i^T XDX^T \mathbf{a}_i + \mathbf{a}_k^T XDX^T \mathbf{a}_k\right)} = \sum_{i=1}^{k-1} \mu_i \mathbf{a}_{k-1}^T (XDX^T)^{-1} \mathbf{a}_i
\end{aligned} \tag{33}$$

We denote:

$$\boldsymbol{\mu}_{k-1} = [\mu_1, \ldots, \mu_{k-1}]^T$$

Then, Eq. 33 can be represented in a single matrix relationship

$$2\frac{A_{k-1}^{T}(XDX^{T})^{-1}XLX^{T}\mathbf{a}_{k}}{tr(A^{T}XDX^{T}A)} = B_{k-1}\boldsymbol{\mu}_{k-1} \tag{34}$$

Thus

$$\boldsymbol{\mu}_{k-1} = 2\frac{(B_{k-1})^{-1}A_{k-1}^{T}(XDX^{T})^{-1}XLX^{T}\mathbf{a}_{k}}{tr(A^{T}XDX^{T}A)} \tag{35}$$

Substituting Eq. 35 into Eq. 29 and combining Eq. 30, we have

$$\frac{2XLX^{T}\mathbf{a}_{k}}{tr(A^{T}XDX^{T}A)} - \lambda\frac{2XDX^{T}\mathbf{a}_{k}}{tr(A^{T}XDX^{T}A)} + \frac{2A_{k-1}(B_{k-1})^{-1}A_{k-1}^{T}(XDX^{T})^{-1}XLX^{T}\mathbf{a}_{k}}{tr(A^{T}XDX^{T}A)} = 0 \tag{36}$$

Then we can obtain

$$XLX^{T}\mathbf{a}_{k} + A_{k-1}(B_{k-1})^{-1}A_{k-1}^{T}(XDX^{T})^{-1}XLX^{T}\mathbf{a}_{k} = \lambda XDX^{T}\mathbf{a}_{k} \tag{37}$$

and

$$\left\{I - (XDX^{T})^{-1}A_{k-1}(B_{k-1})^{-1}A_{k-1}^{T}\right\} \cdot (XDX^{T})^{-1}XLX^{T}\mathbf{a}_{k} = \lambda\mathbf{a}_{k} \tag{38}$$

□

From Theorem 3 we can know, if we use a step-by-step method to get the orthogonal vectors of trace ratio based orthogonal LPP, OLPP is equivalent to trace ratio based orthogonal LPP. That is, OLPP, which suffers from the heavy computation problem since it uses a step-by-step procedure to solve each discriminant vector and involves computing the inverse of matrices, is somewhat like a trace ratio problem. However, our proposed OCDLPP, which is a ratio trace problem, computes the set of orthogonal discriminant vectors simultaneously.

4 Experiments and Results

In this section, experiments are conducted on four well-known face image databases, i.e. ORL, Yale, FERET and PIE, to evaluate the performance of the proposed OCDLPP algorithm. A nearest neighbor classifier with cosine distance is employed to classify in the projected feature space.

4.1 Database

The ORL, Yale, FERET and PIE face databases are used in our experiments. The ORL face database (http://www.cam-orl.co.uk/facedatabase.html) consists of a total of 400 face images, of a total of 40 people (10 samples per person). For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glassed/no glassed). All the images were taken against a dark homogeneous background with the subjects in an upright, front position (with tolerance for some side movement). We have performed some pre-processing on the raw images to align the eyes and eliminate the non-relevant background. In our experiments, each image in ORL database was finally resized to 32 × 32.

Fig. 1 Images of one person in ORL

Fig. 2 Images of one person in Yale

Fig. 3 Images of one person in FERET

Fig. 4 Images of one person in PIE

The Yale face database (http://cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 gray scale images of 15 individuals, each individual has 11 images. The images demonstrate variations in lighting condition, facial expression (normal, happy, sad, sleepy, surprised, and wink). We have performed some pre-processing on the raw images to align the eyes and eliminate the non-relevant background. In our experiments, each image in Yale database was finally resized to 32×32.

The FERET face database, which was sponsored by the Department of Defense of US through the DARPA Program (http://www.itl.nist.gov/iad/humanid/feret/feret_master.html), contains 14,126 images from 1,199 individuals. In our experiments, we select a subset which contains 1,400 images of 200 individuals (each individual has seven images). The subset involves variations in facial expression, illumination and pose. In our experiments, the facial portion of each original was automatically cropped based on the location of eyes and the cropped image was resized to 80×80 pixels.

The CMU PIE face database (http://www.ri.cmu.edu/projects/project418.html) contains 68 individual with 41,368 face images as a whole. The face images were captured under varying pose, illumination, and expression. In our experiments, we select a subset (C29) which contains 1,632 images of 68 individuals (each individual has twenty-four images). The C29 subset involves variations in illumination, facial expression and pose. All of these face images are aligned based on eye coordinates and cropped to 64×64.

Figures 1, 2, 3 and 4 show the sample images from the four databases.

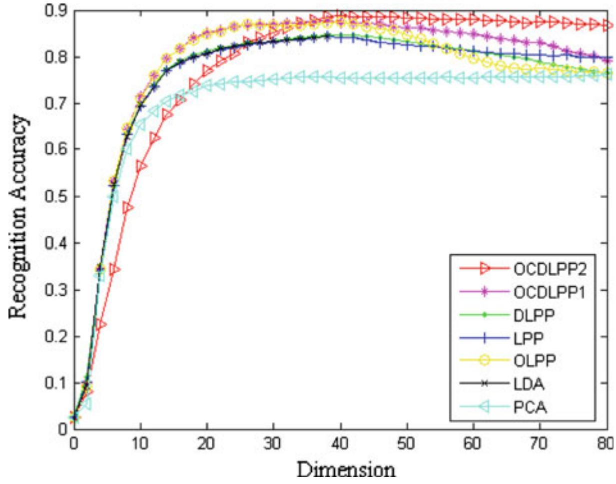

Fig. 5 Recognition accuracy versus dimensionality reduction on the ORL database (3 train)

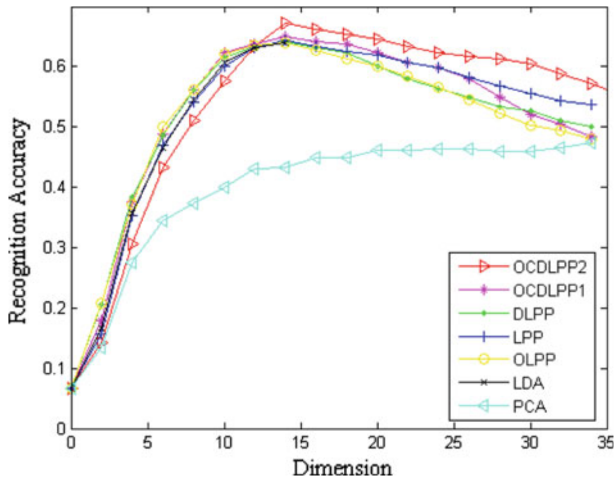

Fig. 6 Recognition accuracy versus dimensionality reduction on the Yale database (3 train)

4.2 Experimental Results and Analysis

In our experiments, the performances of different algorithms, i.e. PCA [3], LDA [4], LPP [26], OLPP [36], DLPP [27] and the proposed method, are tested. For LPP, OLPP, DLPP and OCDLPP algorithms, the Gaussian Kernel $\exp(-\|x-y\|^2/\sigma^2)$ is used and parameter σ is set as $2^{(e-10)/2.5}\delta_0$, $e = 0, 1, \ldots, 20$, where δ_0 is the standard derivation of the training data set. LDA, LPP, OLPP and DLPP involve a preceding PCA stage to avoid the singularity problem and 98% image energy is kept in PCA phase. Note that in our experiments, there are two kinds of OCDLPP algorithms, i.e. OCDLPP1 and OCDLPP2. OCDLPP1 also involves a preceding PCA stage and 98% image energy is kept in PCA phase, while OCDLPP2 does not involve a PCA stage.

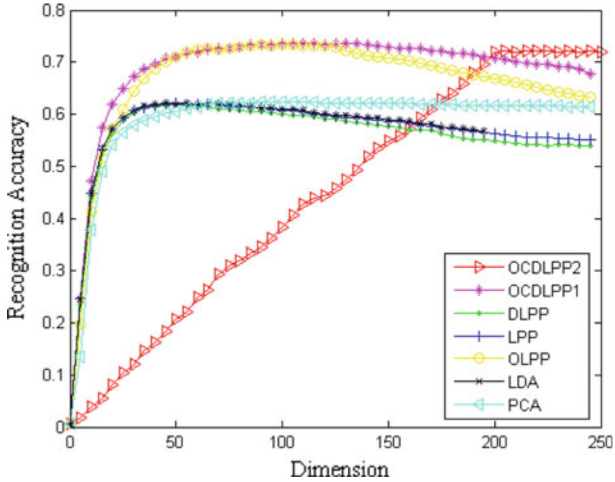

Fig. 7 Recognition accuracy versus dimensionality reduction on the FERET database (3 train)

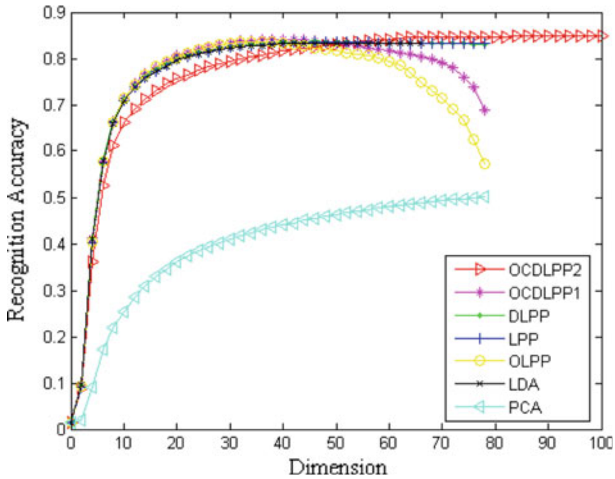

Fig. 8 Recognition accuracy versus dimensionality reduction on the PIE database (3 train)

In the first experiment, we investigate the performances of different algorithms over the reduced dimensions. Three face images of each individual are selected randomly for training while the rest image are used for testing on ORL, Yale, FERET and PIE face databases. We performed 20 times to randomly choose the training set. The final result is the average recognition rate over 20 random training sets. Note that, in terms of theoretical analysis, an upper bound of the dimensionality of the reduced space is $C - 1$ for LDA, where C is the number of classes. Figures 5, 6, 7 and 8 demonstrate the recognition rates of different algorithms over the variance of the dimensionality of subspaces.

In the second experiment, we investigate the performances of different algorithms over different sizes of the training dataset. We randomly select i ($i = 2, 3, 4$ for ORL, Yale and PIE and $i = 2, 3$ for FERET) samples of each individual for training, and the remaining ones are used for testing. The recognition accuracies of different algorithms on ORL, Yale, FERET

Table 1 Recognition accuracy (%) on ORL (mean ± std)

Size	PCA	LDA	LPP	OLPP	DLPP	OCDLPP1	OCDLPP2
2	66.4 ± 3.4	75.4 ± 3.1	75.9 ± 3.1	78.3 ± 3.0	77.6 ± 3.4	78.5 ± 3.2	**80.5 ± 3.6**
3	76.7 ± 2.4	85.1 ± 1.9	86.0 ± 2.1	88.3 ± 1.9	85.7 ± 2.0	88.6 ± 1.8	**89.7 ± 1.8**
4	82.0 ± 2.1	91.3 ± 1.9	92.7 ± 1.7	93.9 ± 1.9	91.7 ± 1.8	94.4 ± 1.5	**94.7 ± 1.6**

The bold values indicate that the corresponding methods obtain the best performances under specific conditions

Table 2 Recognition accuracy (%) on Yale (mean ± std)

Size	PCA	LDA	LPP	OLPP	DLPP	OCDLPP1	OCDLPP2
2	42.9 ± 3.4	52.1 ± 5.6	53.2 ± 5.6	51.6 ± 5.2	52.9 ± 5.7	52.9 ± 5.6	**56.4 ± 4.4**
3	48.8 ± 4.0	65.4 ± 4.6	65.6 ± 4.7	66.1 ± 4.4	66.1 ± 4.8	67.2 ± 5.5	**69.0 ± 4.6**
4	53.8 ± 4.8	72.1 ± 5.4	74.3 ± 5.4	75.5 ± 4.0	74.1 ± 4.8	76.3 ± 5.0	**77.5 ± 5.1**

The bold values indicate that the corresponding methods obtain the best performances under specific conditions

Table 3 Recognition accuracy (%) on FERET (mean ± std)

Size	PCA	LDA	LPP	OLPP	DLPP	OCDLPP1	OCDLPP2
2	47.4 ± 8.7	42.9 ± 7.4	46.3 ± 6.8	54.6 ± 6.5	43.0 ± 6.9	57.7 ± 6.5	**58.2 ± 6.5**
3	62.9 ± 7.8	63.0 ± 7.9	63.3 ± 7.8	75.2 ± 6.4	62.3 ± 7.1	75.3 ± 6.6	**75.4 ± 6.4**

The bold values indicate that the corresponding methods obtain the best performances under specific conditions

and PIE databases are reported on the Tables 1, 2, 3 and 4 respectively. For each i, we average the results over 20 random splits and report the mean as well as the standard deviation.

From Figures 5, 6, 7 and 8 and Tables 1, 2, 3 and 4, we can obtain the following conclusions:

(1) Our proposed OCDLPP consistently outperforms PCA, LDA, LPP, OLPP and DLPP. The reason may be that OCDLPP is implemented without loss of discriminatory information and the transformation matrix is orthogonal.
(2) In Table 3, PCA outperforms LDA when the sample size is two. This is consistent with the observation in [37] that Eigenface can outperform Fisherface when the training set is small.
(3) In Fig. 7 of the experimental results on FERET, we have an interesting observation that OCDLPP2 produces worse result than other algorithms when the subspace dimension is low. The reason may be that the number of classes C on FERET is much larger than those on other databases. When the number of classes C is large, other algorithms achieve the best classification performance in a subspace lower than $C-1$, while OCDLPP2 achieves the best result in a subspace higher than $C-1$.

5 Conclusions

In this paper, we have proposed a novel orthogonal complete discriminant locality preserving projections (OCDLPP) method for face recognition. All training samples are projected into the range of S_t^L to reduce dimensionality without loss on discriminative information. The

Table 4 Recognition accuracy (%) on PIE (mean ± std)

Size	PCA	LDA	LPP	OLPP	DLPP	OCDLPP1	OCDLPP2
2	39.3 ± 1.5	76.6 ± 1.8	76.8 ± 1.7	75.6 ± 2.1	76.8 ± 1.8	76.8 ± 1.8	**78.3** ± **2.0**
3	50.3 ± 1.7	83.9 ± 1.3	84.0 ± 1.3	84.0 ± 1.1	84.1 ± 1.3	84.7 ± 1.1	**86.1** ± **1.2**
4	58.8 ± 2.0	87.6 ± 1.1	87.7 ± 0.9	87.7 ± 1.2	87.8 ± 1.0	88.3 ± 1.1	**89.6** ± **1.0**

The bold values indicate that the corresponding methods obtain the best performances under specific conditions

final optimal discriminant transformation of OCDLPP is orthogonal and is obtained simultaneously. Experimental results on ORL, Yale, FERET and PIE face databases indicate that OCDLPP performs significantly better than DLPP, OLPP, LPP, LDA and PCA methods in terms of recognition accuracy.

Acknowledgments This research is supported by NSFC of China (No. 60632050, 60705006, 60873151, 60973098) and the 2010 Graduates' Research Innovation Program of Higher Education of Jiangsu Province (No. 178). The authors would like to thank the anonymous reviews and the editor for their helpful comments and suggestions to improve the quality of this paper.

References

1. Chellappa R, Wilson C, Sirohey S (1995) Human and machine recognition of faces: a survey. Proc IEEE 83(5):705–740
2. Zhao W, Chellappa R, Phillips P, Rosenfeld A (2003) Face recognition: a literature survey. ACM Comput Surv 35(4):399–458
3. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York
4. Belhumeour PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):11–720
5. Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84(405):165–175
6. Chen LF, Liao HYM, Ko MT, Yu GJ (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognit 33(1):1713–1726
7. Huang R, Liu Q, Lu H, Ma S (2002) Solving the small size problem of LDA. In: Proceedings of the 16th international conference on pattern recognition, pp 29–32
8. Ye J (2005) Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. J Mach Learn Res 6:483–502
9. Park H, Drake B, Lee S, Park C (2007) Fast linear discriminant analysis using QR decomposition and regularization. Technical Report GT-CSE-07-21
10. Howland P, Park H (2004) Generating discriminant analysis using the generalized singular value decomposition. IEEE Trans Pattern Anal Mach Intell 26(8):995–1006
11. Ye J, Janardan R, Park CH, Park H (2004) An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE Trans Pattern Anal Mach Intell 26(8):982–994
12. Howland P, Jeon M, Park H (2003) Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. SIAM J Matrix Anal Appl 25(1):165–179
13. Tao D, Li X, Wu X, Maybank SJ (2009) Geometric mean for subspace selection. IEEE Trans Pattern Anal Mach Intell 31(2):260–274
14. Bian W, Tao D (2008) Harmonic mean for subspace selection. In: 19th International conference on pattern recognition, 2008. ICPR 2008, Tampa, FL, pp 1–4
15. Tenenbaum JB, Silva Vd, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
16. Roweis ST, Saul LK (2000) Nonlinear dimension reduction by locally linear embedding. Science 290:2323–2326
17. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396
18. Yan S, Xu D, Zhang B, Zhang H-J, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51

19. Zhang T, Tao D, Li X, Yang J (2009) Patch alignment for dimensionality reduction. IEEE Trans Knowl Data Eng 21(9):1299–1313
20. Zhou T, Tao D (2009) Manifold elastic net for sparse learning. In: Proceedings of the 2009 IEEE international conference on systems, man, and cybernetics, San Antonio, TX, USA, pp 3699–3704
21. Zhou T, Tao D, Wu X (2010) Manifold elastic net: a unified framework for sparse dimension reduction. Data Min Knowl Discov. doi:10.1007/s10618-010-0182-x
22. Geng B, Xu C, Tao D, Yang L, Hua X-S (2009) Ensemble manifold regularization. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009, Miami, FL, pp 2396–2402
23. Bian W, Tao D (2009) Manifold regularization for SIR with rate root-n convergence. In: Advances in neural information processing systems (NIPS), pp 1–8
24. Si S, Tao D, Geng B (2010) Bregman divergence-based regularization for transfer subspace learning. IEEE Trans Knowl Data Eng 22(7):929–942
25. He X, Yan S, Hu Y, Zhang H (2003) Learning a locality preserving subspace for visual recognition. In: Proceedings of the ninth international conference on computer vision, France, pp 385–392
26. He X, Yan S, Hu Y, Niyogi P, Zhang H (2005) Face recognition using Laplacian faces. IEEE Trans Pattern Anal Mach Intell 27(3):328–340
27. Yu W, Teng X, Liu C (2006) Face recognition using discriminant locality preserving projections. Image Vis Comput 24:239–248
28. Yang L, Gong W, Gu X, Li W, Liang Y (2008) Null space discriminant locality preserving projections for face recognition. Neurocomputing 71:3644–3649
29. Hu H (2008) Orthogonal neighborhood preserving discriminant analysis for face recognition. Pattern Recognit 412:2045–2054
30. Wang H, Yan S, Xu D, Tang X, Huang TS (2007) Trace ratio vs. ratio trace for dimensionality reduction. In: IEEE computer society conference on computer vision and pattern recognition (CVPR'07)
31. Foley DH, Sammon JWJ (1975) An optimal set of discriminant vectors. IEEE Trans Comput 24(3): 281–289
32. Duchene J, Leclercq S (1988) An optimal transformation for discriminant and principal component analysis. IEEE Trans Pattern Anal Mach Intell 10(6):978–983
33. Ma W-Y (2005) OCFS: optimal orthogonal centroid feature selection for text categorization. In: Proceedings of the 28th international ACM SIGIR conference on research and development in information retrieval, New York, pp 122–129
34. Wang H, Chen S, Hu Z, Zheng W (2008) Locality-preserved maximum information projection. IEEE Trans Neural Netw 19(4):571–585
35. Kokiopoulou E, Saad Y (2007) Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique. IEEE Trans Pattern Anal Mach Intell 29(2):2143–2156
36. Cai D, He X, Han J, Zhang H-J (2006) Orthogonal laplacianfaces for face recognition. IEEE Trans Image Process 15(11):3608–3614
37. Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233