Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 1156–1163
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Neighborhood discriminant projection for face recognition

Qubo You *, Nanning Zheng, Shaoyi Du, Yang Wu

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi Province 710049, PR China

Received 12 February 2006; received in revised form 19 December 2006
Available online 4 February 2007

Communicated by R.P.W. Duin

Abstract

We propose a novel manifold learning approach, called *Neighborhood Discriminant Projection* (NDP), for robust face recognition. The purpose of NDP is to preserve the within-class neighboring geometry of the image space, while keeping away the projected vectors of the samples of different classes. For representing the intrinsic within-class neighboring geometry and the similarity of the samples of different classes, the *within-class affinity weight* and the *between-class affinity weight* are used to model the *within-class submanifold* and the *between-class submanifold* of the samples, respectively. Comprehensive comparisons and extensive experiments on face recognition are performed to demonstrate the effectiveness and robustness of our proposed method.

Keywords: Face recognition; Manifold learning; Within-class submanifold; Between-class submanifold; Linear subspace

1. Introduction

Face recognition has become one of the most challenging problems in computer vision and pattern recognition (Chellappa et al., 1995; Zhao et al., 2003; Li and Jain, 2005). Numerous methods have been proposed for face recognition over the past few decades (Zhao et al., 2003; Turk and Pentland, 1991; Belhumeur et al., 1997; Cevikalp et al., 2005; Huang et al., 2002). Among these methods, Principal Component Analysis (PCA) (Turk and Pentland, 1991) and Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997) are the most popular techniques, which assume that the samples lie on a linearly embedded manifold and aim at preserving the global Euclidean structure of the image space.

However, a lot of research has shown that facial images possibly lie on a nonlinear submanifold (Roweis and Saul, 2000; Tenenbaum et al., 2000; He and Niyogi, 2003; He et al., 2005b; Chen et al., 2005; Hu et al., 2004; Yang, 2002b). When using PCA and LDA for dimensionality reduction, they will fail to discover the intrinsic dimension of the image space. Recently, a number of nonlinear techniques originated from manifold learning have been proposed to discover the nonlinear structure of the manifold by investigating the local geometry of samples, such as LLE (Roweis and Saul, 2000), Isomap (Tenenbaum et al., 2000) and Laplacian Eigenmap (Belkin and Niyogi, 2001). These methods are appropriate for representation of nonlinear data, but only defined on the training data. Because of the difficult issue that how to map a new test sample to the low dimensional space, these nonlinear manifold learning algorithms cannot be applied directly to classification problems. Although kernel-based methods, such as Kernel PCA (Schölkopf et al., 1998) and Kernel LDA (Yang, 2002a), can efficiently deal with the nonlinear data and evaluate the map on new test samples, they do not consider explicitly the structure of the nonlinear manifold of the data and are computationally expensive in computational complexity which is undesirable in practical face recognition systems.

* Corresponding author. Tel.: +86 029 82668802x8002; fax: +86 029 82668672.
E-mail address: youqubo@gmail.com (Q. You).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.01.011

Recently, some manifold-based algorithms (He and Niyogi, 2003; He et al., 2003, 2005a,b; Athitsos et al., 2004) resolve the difficulty that how to map a new test sample to the low dimensional space. However, these methods are designed to preserve the locality of samples in the lower dimensional space rather than good discrimination ability. As a result, the projected vectors of different classes may overlap. Only a few manifold learning algorithms thoroughly consider the within-class information and between-class information to address classification problems, including (Yan et al., 2004, 2005; Chen et al., 2005; Yang, 2002b). Among these methods, Local Discriminant Embedding (LDE) (Chen et al., 2005) and Marginal Fisher Analysis (MFA) (Yan et al., 2005) are the representational methods. However, LDE and MFA both use PCA for dimensionality reduction to deal with the small sample size problem, which may result in the loss of important discriminative information.

In order to overcome the above shortcoming, we propose a novel manifold learning algorithm, called *Neighborhood Discriminant Projection* (NDP), which explicitly considers the *within-class submanifold* and the *between-class submanifold* by integrating the within-class neighboring information and the between-class neighboring information. The aim of NDP is to preserve the within-class neighboring geometry of the image space, while keeping away the projected vectors of the samples of different classes. To be specific, the within-class submanifold is modeled by the *within-class affinity weight* of the samples, which is an optimal representation of the intrinsic neighboring geometry and computed based on the method of LLE (Roweis and Saul, 2000). The between-class submanifold is modeled by the *between-class affinity weight* of the samples, which reflects the similarity of the samples of different classes and can be obtained by the method of Laplacian Eigenmap (Belkin and Niyogi, 2001). Due to thoroughly consider the nonlinear within-class submanifold and the between-class submanifold, the obtained linear subspace not only preserves the within-class neighboring geometry but also differentiates the projected vectors of the samples of different classes.

The rest of the paper is organized as follows: the Neighborhood Discriminant Projection (NDP) algorithm is described in Section 2. In Section 3, experimental results are presented to demonstrate the effectiveness and robustness of NDP. Finally, conclusions are summarized in Section 4.

2. Neighborhood discriminant projection (NDP)

For the convenience of understanding, in the following, the small *italic* letters denote scalars, such as a, b, c; the small **bold** non-italic letters denote vectors, such as $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$; and the capital **bold** non-italic letters denote matrices, such as $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Let we have n samples $\{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ belonging to c classes of faces, and the corresponding class labels are $\{t_i \mid t_i \in \{1, 2, \ldots, c\}\}_{i=1}^n$. And let the number of samples in the ith class be n_i, while $\mathbf{x}_j^i$ denotes the jth sample in the ith class.

2.1. Modeling within-class submanifold

Recall the LLE algorithm (Roweis and Saul, 2000), LLE supposes that the data point, sampling from the manifold of the data, can be reconstructed by a linear combination of its k-nearest neighbors; furthermore, the locally geometric characteristic is valid for the local neighbors on the manifold of the data. As a result, the low dimensional embedding of LLE preserves the neighboring geometry of the high dimensional space (Saul and Roweis, 2003). One of the aims of NDP is to preserve the within-class neighboring geometry. Therefore, similar to LLE, we assume that every facial image $\mathbf{x}_j^i$, sampling from the nonlinear submanifold of the image space, can be reconstructed by the linear combination of the other samples in the ith class. Moreover, the weight w_{ij} , reflecting the contribution of the jth facial image to the reconstruction of the ith facial image, should be preserved in the lower dimensional facial subspace. We call w_{ij} the *within-class affinity weight*. In order to compute w_{ij}, we can first construct the *within-class graph* $\mathbf{G}$, where the ith node corresponds to the sample $\mathbf{x}_i$. When the pair of samples $\mathbf{x}_i$ and $\mathbf{x}_j$ from the same class, i.e. $t_i = t_j$, an edge is added on the within-class graph $\mathbf{G}$. The within-class affinity weight matrix $\mathbf{W}$ can be computed by minimizing the following reconstruction error (Saul and Roweis, 2003):

$$\min \sum_i \left\| \mathbf{x}_i - \sum_j w_{ij} \mathbf{x}_j \right\|^2 \tag{1}$$

With two constraints (Saul and Roweis, 2003):

(1) $w_{ij} = 0$, if there is no edge from the ith node to the jth node in the within-class graph $\mathbf{G}$;
(2) $\sum_j w_{ij} = 1, i = 1, 2, \ldots, n$;

Considering some sample $\bar{\mathbf{x}}^i$ in the ith class and other samples $\bar{\mathbf{x}}_j^i$ in the ith class, $\sum_j \bar{w}_j \bar{\mathbf{x}}_j^i$ is used to reconstruct $\bar{\mathbf{x}}^i$. According to LLE

$$\bar{w}_j = \frac{\sum_k h'_{jk}}{\sum_{lm} h'_{lm}}, \quad \mathbf{H}' = \mathbf{H}^{-1}, \quad h_{jk} = (\bar{\mathbf{x}}^i - \bar{\mathbf{x}}_j^i)^{\mathrm{T}} (\bar{\mathbf{x}}^i - \bar{\mathbf{x}}_k^i) \tag{2}$$

With the two constraints, the weight w_{ij} is invariant to rotation, rescaling, and translation (Saul and Roweis, 2003). Therefore, the within-class affinity weight matrix $\mathbf{W}$ preserves the intrinsic local geometry of the samples in the same class.

Let $\mathbf{Q} \in \mathbb{R}^{d \times \ell}$ be the transformation matrix, and $\{\mathbf{y}_i = \mathbf{Q}^{\mathrm{T}} \mathbf{x}_i \mid \mathbf{y}_i \in \mathbb{R}^\ell\}_{i=1}^n$ are projected vectors of the facial images $\{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$. In order to make the projected vectors preserve the local geometry of the image space, according to LLE (Saul and Roweis, 2003), we should minimize the following cost function:

$$
\begin{aligned}
J_{\min}(\mathbf{Q}) &= \sum_i \left\| \mathbf{y}_i - \sum_j w_{ij}\mathbf{y}_j \right\|^2 = \sum_i \left\| \mathbf{Q}^{\mathrm{T}}\mathbf{x}_i - \sum_j w_{ij}(\mathbf{Q}^{\mathrm{T}}\mathbf{x}_j) \right\|^2 \\
&= \sum_i \mathrm{tr}\left\{ \left(\mathbf{Q}^{\mathrm{T}}\mathbf{x}_i - \sum_j w_{ij}(\mathbf{Q}^{\mathrm{T}}\mathbf{x}_j)\right)\left(\mathbf{Q}^{\mathrm{T}}\mathbf{x}_i - \sum_j w_{ij}(\mathbf{Q}^{\mathrm{T}}\mathbf{x}_j)\right)^{\mathrm{T}} \right\} \\
&= \mathrm{tr}\left\{ \mathbf{Q}^{\mathrm{T}}\left(\sum_i \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{Q} \right\} - \mathrm{tr}\left\{ \mathbf{Q}^{\mathrm{T}}\left(\sum_i\sum_j \mathbf{x}_i w_{ij}\mathbf{x}_j^{\mathrm{T}}\right)\mathbf{Q} \right\} \\
&\quad - \mathrm{tr}\left\{ \mathbf{Q}^{\mathrm{T}}\left(\sum_i\sum_j \mathbf{x}_j w_{ij}\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{Q} \right\} \\
&\quad + \mathrm{tr}\left\{ \mathbf{Q}^{\mathrm{T}}\left(\sum_i \left(\sum_j w_{ij}\mathbf{x}_j\right)\left(\sum_j w_{ij}\mathbf{x}_j^{\mathrm{T}}\right)\right)\mathbf{Q} \right\} \\
&= \mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{X}(\mathbf{I}-\mathbf{W})^{\mathrm{T}}(\mathbf{I}-\mathbf{W})\mathbf{X}^{\mathrm{T}}\mathbf{Q}) \qquad (3)
\end{aligned}
$$

where the symbol "tr" denotes the operation of trace, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, and $\mathbf{I} = \mathrm{diag}(1, 1, \ldots, 1)$.

2.2. Modeling between-class submanifold

Since the purpose of NDP is to solve the classification problems, we should make the projected vectors of the samples of different classes far from each other. In order to make use of the between-class neighboring information of all the samples, we can construct the *between-class graph* $\mathbf{G}'$, where the ith node corresponds to the sample $\mathbf{x}_i$. Considering the pair of samples $\mathbf{x}_i$ and $\mathbf{x}_j$ from different classes, i.e. $t_i \neq t_j$, an edge is added on the between-class graph $\mathbf{G}'$ if $\mathbf{x}_i$ is one of $\mathbf{x}_j$'s k-nearest neighbors or $\mathbf{x}_j$ is one of $\mathbf{x}_i$'s k-nearest neighbors. Let weight w'_{ij} reflect the weight of the edge from the ith node to the jth node in the between-class graph $\mathbf{G}'$, and is defined as heat kernel (Belkin and Niyogi, 2001).

$$
w'_{ij} = \begin{cases} \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t) \\ \quad \text{if } i\text{th node and } j\text{th node are connected} \\ 0 \quad \text{otherwise} \end{cases} \qquad (4)
$$

We call w'_{ij} the *between-class affinity weight*. If the pair of facial images $\mathbf{x}_i$ and $\mathbf{x}_j$ from the different classes is distant, the weight w'_{ij} will be very small or further $w'_{ij} = 0$. Therefore, the weight w'_{ij} reinforces the pair of facial images $\mathbf{x}_i$ and $\mathbf{x}_j$ from different classes if $\mathbf{x}_i$ is one of $\mathbf{x}_j$'s k-nearest neighbors or $\mathbf{x}_j$ is one of $\mathbf{x}_i$'s k-nearest neighbors. In order to make the projected vectors of the samples of different classes far from each other, we can maximize the following cost function:

$$
\begin{aligned}
J_{\max}(\mathbf{Q}) &= \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w'_{ij} = \sum_{ij} \|\mathbf{Q}^{\mathrm{T}}\mathbf{x}_i - \mathbf{Q}^{\mathrm{T}}\mathbf{x}_j\|^2 w'_{ij} \\
&= \sum_{ij} \mathrm{tr}\{\mathbf{Q}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{Q}\} w'_{ij} \\
&= \mathrm{tr}\left\{ \sum_{ij} \mathbf{Q}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j) w'_{ij} (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{Q} \right\} \\
&= \mathrm{tr}\left\{ \mathbf{Q}^{\mathrm{T}} \sum_{ij} (\mathbf{x}_i - \mathbf{x}_j) w'_{ij} (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{Q} \right\} \\
&= \mathrm{tr}\left\{ \mathbf{Q}^{\mathrm{T}} \left(\sum_{ij} \mathbf{x}_i w'_{ij}\mathbf{x}_i^{\mathrm{T}} + \sum_{ij} \mathbf{x}_j w'_{ij}\mathbf{x}_j^{\mathrm{T}} - \sum_{ij} \mathbf{x}_i w'_{ij}\mathbf{x}_j^{\mathrm{T}} - \sum_{ij} \mathbf{x}_j w'_{ij}\mathbf{x}_i^{\mathrm{T}} \right)\mathbf{Q} \right\} \qquad (5)
\end{aligned}
$$

Since the matrix $\mathbf{W}'$ is symmetric, we can rewrite Eq. (5) as

$$
\begin{aligned}
J_{\max}(\mathbf{Q}) &= \mathrm{tr}\{\mathbf{Q}^{\mathrm{T}}(2\mathbf{X}\mathbf{D}'\mathbf{X}^{\mathrm{T}} - 2\mathbf{X}\mathbf{W}'\mathbf{X}^{\mathrm{T}})\mathbf{Q}\} \\
&= 2\mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{X}(\mathbf{D}' - \mathbf{W}')\mathbf{X}^{\mathrm{T}}\mathbf{Q}) \\
&\propto \mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{X}(\mathbf{D}' - \mathbf{W}')\mathbf{X}^{\mathrm{T}}\mathbf{Q}) \qquad (6)
\end{aligned}
$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ and $\mathbf{D}'$ is diagonal matrix with diagonal element $d'_{ii} = \sum_j w'_{ij}$.

2.3. Low dimensional embedding

From the above theoretic analysis, we should not only maximize the $J_{\max}(\mathbf{Q})$ but also minimize the $J_{\min}(\mathbf{Q})$, i.e. maximizing the following criterion:

$$
J_{\mathrm{NDP}}(\mathbf{Q}_{\mathrm{opt}}) = \arg\max_{\mathbf{Q}} \frac{\mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_{\mathrm{BN}}\mathbf{Q})}{\mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_{\mathrm{WN}}\mathbf{Q})} \qquad (7)
$$

where $\mathbf{S}_{\mathrm{BN}} = \mathbf{X}(\mathbf{D}' - \mathbf{W}')\mathbf{X}^{\mathrm{T}}$ and $\mathbf{S}_{\mathrm{WN}} = \mathbf{X}(\mathbf{I} - \mathbf{W})^{\mathrm{T}}(\mathbf{I} - \mathbf{W})\mathbf{X}^{\mathrm{T}}$. The matrix $\mathbf{S}_{\mathrm{BN}}$ is called the *between-class neighborhood scatter matrix*. The matrix $\mathbf{S}_{\mathrm{WN}}$ is called the *within-class neighborhood scatter matrix*. The rank of the matrix $\mathbf{S}_{\mathrm{WN}}$ is at most $n - c$, while the size of the matrix $\mathbf{S}_{\mathrm{WN}}$ is $d \times d$. Due to the small sample size problem $d \gg n$ (Fukunaga, 1990), however, the matrix $\mathbf{S}_{\mathrm{WN}}$ is singular and cannot be applied directly to compute transformation matrix $\mathbf{Q}$ based on Eq. (7).

Recently, Cevikalp et al. (2005) has used Discriminative Common Vector method (DCV), a variation of LDA, to resolve the small sample size problem. In this method, the lower linear subspace, where all the training samples in the same class correspond to a unique discriminative common vector, is obtained from the null space of the within-class scatter matrix in LDA. However, DCV aims at preserving the global Euclidean structure of the image space and assumes that the samples lie on a linearly embedded manifold. Instead of using the within-class scatter matrix, we use the within-class neighborhood scatter matrix to compute the null space, where the within-class neighboring geometry is preserved, and then extract the lower linear facial subspace where the projected vectors of the samples of different classes are far from each other.

In a special case, where $\mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_{\mathrm{WN}}\mathbf{Q}) = 0$, the criterion in Eq. (7) can be rewritten as

$$
J_{\mathrm{NDP}}(\mathbf{Q}_{\mathrm{opt}}) = \arg\max_{\substack{\mathbf{Q} \\ \mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_{\mathrm{WN}}\mathbf{Q})=0}} \mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_{\mathrm{BN}}\mathbf{Q}) \qquad (8)
$$

Since $\mathbf{S}_{\mathrm{WN}}$ is symmetric positive semi-definite, $\mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_{\mathrm{WN}}\mathbf{Q}) = 0$ means that the column vectors of transformation matrix $\mathbf{Q}$ are the eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_{\mathrm{WN}}$. Therefore, we can project the training samples onto the null space of the $\mathbf{S}_{\mathrm{WN}}$ and then compute the projection directions by maximizing $\mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}_{\mathrm{BN}}\mathbf{Q})$. As a result, we must first obtain the eigenvectors which span the null space of $\mathbf{S}_{\mathrm{WN}}$. Due to the small sample size problem, directly computing the bases spanning the null space of $\mathbf{S}_{\mathrm{WN}}$ is intractable. However, we

can first compute the bases spanning the range space of $\mathbf{S}_{\mathrm{WN}}$ and then compute the null space of $\mathbf{S}_{\mathrm{WN}}$.

Suppose $\mathrm{rank}(\mathbf{S}_{\mathrm{WN}}) = r$, B be the null space of $\mathbf{S}_{\mathrm{WN}}$ and $B^{\perp}$ be the range space of $\mathbf{S}_{\mathrm{WN}}$, where

$$B = \mathrm{span}\{\gamma_k | \gamma_k \mathbf{S}_{\mathrm{WN}} = 0, k = 1, \ldots, d-r\} \tag{9}$$

$$B^{\perp} = \mathrm{span}\{\gamma_k | \gamma_k \mathbf{S}_{\mathrm{WN}} \neq 0, k = d-r+1, \cdots, d\} \tag{10}$$

Let $\mathbf{U} = [\gamma_1, \ldots, \gamma_{d-r}]$ and $\bar{\mathbf{U}} = [\gamma_{d-r+1}, \ldots, \gamma_d]$, due to $\mathbf{U}\mathbf{U}^{\mathrm{T}} + \bar{\mathbf{U}}\bar{\mathbf{U}}^{\mathrm{T}} = \mathbf{I}$ where $\mathbf{I} = \mathrm{diag}(1, \ldots, 1)$, we can obtain

$$\mathbf{x}_j^i = \mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{x}_j^i + \bar{\mathbf{U}}\bar{\mathbf{U}}^{\mathrm{T}}\mathbf{x}_j^i \tag{11}$$

Let $\mathbf{z}_j^i = \mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{x}_j^i = \sum_{k=1}^{d-r}\langle \gamma_k, \mathbf{x}_j^i\rangle \gamma_k$. Therefore, $\mathbf{z}_j^i$ means that the sample $\mathbf{x}_j^i$ is projected onto the null space of $\mathbf{S}_{\mathrm{WN}}$. According to Eq. (11), we can rewrite $\mathbf{z}_j^i$ as

$$\mathbf{z}_j^i = \mathbf{x}_j^i - \bar{\mathbf{U}}\bar{\mathbf{U}}^{\mathrm{T}}\mathbf{x}_j^i \tag{12}$$

Since the dimension d of facial image is very high (Turk, 2001), computing the eigenvectors corresponding to the non-zero eigenvalues of $\mathbf{S}_{\mathrm{WN}}$ is time-consuming. According to the knowledge of linear algebra, when supposing λ_m and $\mathbf{o}_m$ be the mth non-zero eigenvalue and the corresponding eigenvector of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, $\mathbf{p}_m = (\lambda_m)^{-1/2}\mathbf{A}\mathbf{o}_m$ is the eigenvector, corresponding to the mth non-zero eigenvalue, of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. Let $\mathrm{rank}(\mathbf{A}) = k$, $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_k]$, $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_k]$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, then

$$\mathbf{P} = \mathbf{A}\mathbf{O}\mathbf{\Lambda}^{-1/2} \tag{13}$$

Let $\mathbf{A} = \mathbf{X}(\mathbf{I} - \mathbf{W})^{\mathrm{T}}$, then $\mathbf{S}_{\mathrm{WN}} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. Since the size of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is $n \times n$ ($n \ll d$), we can quickly obtain the $\bar{\mathbf{U}}$ according to Eq. (13). Then we can obtain $\mathbf{z}_j^i$ based on Eq. (12).

After projecting all the training samples onto the null space of $\mathbf{S}_{\mathrm{WN}}$, the criterion $J_{\mathrm{NDP}}(\mathbf{Q}_{\mathrm{opt}})$ of NDP can be represented as

$$J_{\mathrm{NDP}}(\mathbf{Q}_{\mathrm{opt}}) = \arg\max_{\mathbf{Q}} \mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{S}'_{\mathrm{BN}}\mathbf{Q}) \tag{14}$$

where $\mathbf{S}'_{\mathrm{BN}} = \mathbf{Z}(\mathbf{D}' - \mathbf{W}')\mathbf{Z}^{\mathrm{T}}$, $\mathbf{Z} = [\mathbf{Z}^1, \ldots, \mathbf{Z}^c]$ and $\mathbf{Z}^i = [\mathbf{z}_1^i, \ldots, \mathbf{z}_{n_i}^i]$. From Eq. (14), we can find that the projection directions are the leading eigenvectors corresponding to the non-zero eigenvalues of $\mathbf{S}'_{\mathrm{BN}}$. However, computing the eigenvectors corresponding to the non-zero eigenvalues of $\mathbf{S}'_{\mathrm{BN}}$ is also time-consuming. Let λ'_i and $\mathbf{u}'_i$ be the ith eigenvalue and the corresponding eigenvectors of $\mathbf{E}' = \mathbf{D}' - \mathbf{W}' \in \mathbb{R}^{n\times n}$, $\mathbf{V}' = \mathrm{diag}(\lambda'_1, \ldots, \lambda'_n)$ and $\mathbf{U}' = [\mathbf{u}'_1, \ldots, \mathbf{u}'_n]$. According to the knowledge of linear algebra, we can obtain $\mathbf{E}' = \mathbf{D}' - \mathbf{W}' = \mathbf{E}''(\mathbf{E}'')^{\mathrm{T}}$ where $\mathbf{E}'' = \mathbf{U}'(\mathbf{V}')^{1/2}(\mathbf{U}')^{\mathrm{T}}$. Let $\mathbf{F} = \mathbf{Z}\mathbf{E}''$, then $\mathbf{S}'_{\mathrm{BN}} = \mathbf{F}\mathbf{F}^{\mathrm{T}}$. Since the size of $\mathbf{F}^{\mathrm{T}}\mathbf{F}$ is $n \times n$ ($n \ll d$), we can quickly obtain the transformation matrix $\mathbf{Q}$ based on Eq. (13).

Then, the embedding is as follows:

$$\mathbf{y} = \mathbf{Q}^{\mathrm{T}}\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{x} \tag{15}$$

Since the projection direction $\mathbf{q}_i$ belongs to the null space of $\mathbf{S}_{\mathrm{WN}}$, we can represent $\mathbf{q}_i$ as $\mathbf{q}_i = \sum_{j=1}^{d-r}\partial_{i,j}\gamma_j$. Therefore, Eq. (15) can be rewrite as

$$\begin{aligned}\mathbf{y} &= \mathbf{Q}^{\mathrm{T}}\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{x} = [\mathbf{q}_1, \ldots, \mathbf{q}_{c-1}]^{\mathrm{T}}\left(\sum_{k=1}^{d-r}\langle \gamma_k, \mathbf{x}\rangle \gamma_k\right) \\ &= \left[\sum_{k=1}^{d-r}\langle \gamma_k, \mathbf{x}\rangle \partial_{1,k}, \ldots, \sum_{k=1}^{d-r}\langle \gamma_k, \mathbf{x}\rangle \partial_{c-1,k}\right]^{\mathrm{T}} \\ &= [\mathbf{q}_1, \ldots, \mathbf{q}_{c-1}]^{\mathrm{T}}\mathbf{x} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}\end{aligned} \tag{16}$$

Since the projection direction $\mathbf{q}_i$ belongs to the null space of the within-class neighborhood scatter matrix $\mathbf{S}_{\mathrm{WN}}$, the linear facial subspace not only makes the projected vectors of the samples of different classes far from each other but also preserves the within-class neighboring geometric structure, which is more important than the global Euclidean structure in many practical classification problems (He et al., 2005a).

3. Experimental results

To verify the proposed NDP approach, three well-known face databases (ORL database (Samaria and Harter, 1994), UMIST database (Graham and Allinson, 1998) and FERET database (Phillips et al., 2000)) were used; and the system performance of NDP was compared with the ones of PCA (Turk and Pentland, 1991), LDA (Belhumeur et al., 1997), Direct-LDA (Yu and Yang, 2001), DCV (Cevikalp et al., 2005), MFA (Yan et al., 2005), LDE (Chen et al., 2005), Supervised NPE (SNPE) (He et al., 2005b) and Supervised Laplacianfaces (SLF) (He et al., 2005a). For its simplicity, the nearest-neighbor method using Euclidean metric was employed.

3.1. ORL database

In ORL database, there are 10 different grey images for each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). The size of each image is 92×112 pixels with 256 grey levels per pixel. We illustrate the facial images of two individuals from the ORL database in Fig. 1.

In Section 2, we have discussed how to extract a linear facial subspace. Then, the projected vector of a facial image can be obtained from Eq. (16). Similar to PCA and LDA, we can display the projection direction $\mathbf{q}_i$ as an image, called NDP-faces. Using all the facial images in the ORL database as the training set, we illustrate the first 5 projection directions of PCA, LDA and NDP in Fig. 2. It is very interesting to see that the NDP-faces are similar to Fisherfaces.

We chose randomly $\xi(= 3, 4, 5)$ different images per individual to form the training set. The rest of the ORL database was used for testing set. For each given ξ, we performed 50 times to choose randomly the training set.

Fig. 1. Twenty facial images of two individuals in the ORL database.

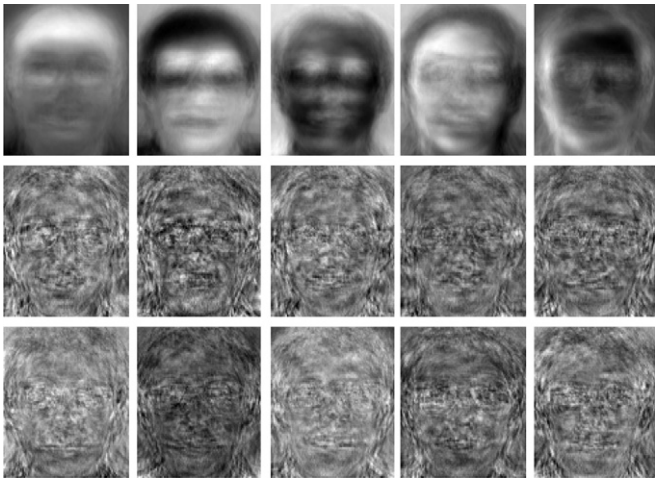

Fig. 2. The first 5 projection directions. From top to bottom: Eigenfaces, Fisherfaces and NDP-faces.

The final result is the average recognition rate over 50 random training sets. In general, the recognition rate changes with the dimension ℓ of reduced space. Fig. 3 illustrates the plot of recognition rate versus the dimension ℓ of reduced space for PCA, LDA, Direct-LDA, DCV, MFA, LDE, SNPE, SLF and NDP. The top recognition rate achieved by each method and the corresponding dimension ℓ of reduced space are also shown in Table 1.

3.2. UMIST database

The UMIST database contains 575 grey images of 20 people. The number of different views per individual varies from 19 to 48. We used a Pre-Cropped version of the UMIST database, where each individual covers a range of views from profile to frontal views. The size of each cropped image is 92×112 pixels with 256 grey levels per pixel. We illustrate the facial images of one individual with 20 different views from the UMIST database in Fig. 4.

In this experiment, we chose randomly three different views per individual to form the training set in the range 0–30°, range 30–60° and range 60–90°, respectively. The rest of the UMIST database was used for testing set. We performed 20 times to choose randomly the training set. The final result is the average recognition rate over 20 training sets selected randomly. Fig. 5 illustrates the plot of recognition rate versus the dimension ℓ of reduced space for PCA, LDA, Direct-LDA, DCV, MFA, LDE, SNPE, SLF and NDP. The top recognition rate achieved by each method and the corresponding dimension ℓ of reduced space are also shown in Table 2.

3.3. FERET database

This dataset consists of all the 1195 people from the FERET Fa/Fb data set. There are two face images for each person. We preprocessed these original images by aligning transformation and scaling transformation so that the two eyes were aligned at the same position. Then, the facial areas were cropped into the resulting images. The size of each cropped image is 64×64, with 256 grey levels per pixel. We did not perform further preprocessing. Fig. 6 shows the 20 facial images of 10 individuals from this dataset.

We selected randomly 495 people for training and used the remaining 700 people as testing. For each testing people, one face image is in the gallery and the other is for probe. We performed 50 times to choose randomly the training set. The final result is the average recognition rate over 50 random training sets. Fig. 7 illustrates the plot of recognition rate versus the dimension ℓ of reduced space for PCA, LDA, Direct-LDA, DCV, MFA, LDE, SNPE, SLF and NDP. The top recognition rate achieved by each method and the corresponding dimension ℓ of reduced space are also shown in Table 3.

3.4. Discussion

From Figs. 3, 5 and 7, it is very obvious that the NDP method outperforms the other methods across all the values of the dimension ℓ of reduced space. From Fig. 3a–c, we can see that the performance of NDP improves significantly as the number ξ of training samples per individual increases.

In practical face recognition system, a probe image is projected onto the facial subspace spanned by projection directions and then compared to the projected vectors of gallery samples. Testing time is used to evaluate the time efficiency of face recognition methods during comparison. From Tables 1–3, we can find that the dimension ℓ of NDP corresponding to the top recognition rate is lower compared with the other methods, that is, the dimension

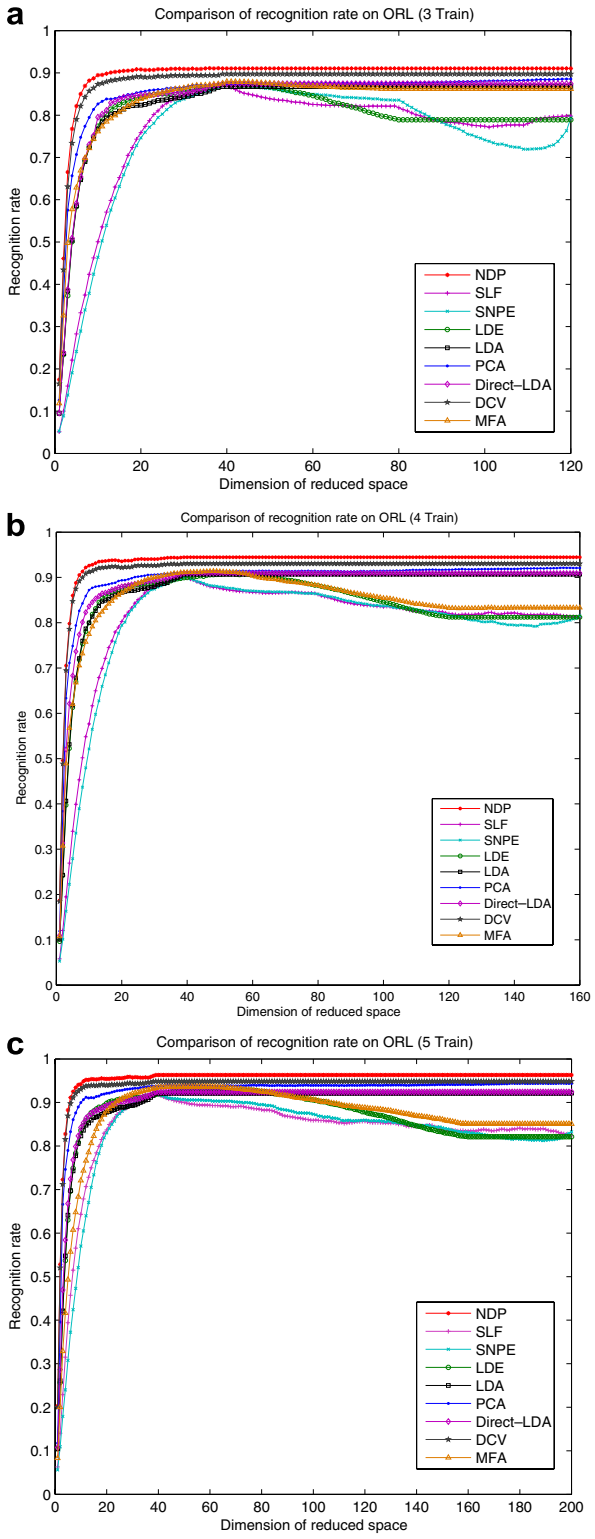

Fig. 3. Recognition rate versus dimension ℓ of reduced space on the ORL database: (a) 3 train, (b) 4 train and (c) 5 train.

of the facial subspace for NDP is lower than the other methods. Thus, the testing time of NDP is less compared with the other methods.

Different from the images in the ORL database with the subjects in an upright, frontal position, each subject in the UMIST database covers a range of views from profile to frontal views. Thus, the result on the UMIST database is more effective to evaluate the performance of each method in the different views. Different from ORL database and UMIST database, where the number of the subject is small, the dataset from the FERET database consists of 1195 people. Moreover, this dataset has no overlap between the training set and gallery/probe set, which results in the requirement of generalizable ability from known subjects in the training set to unknown subjects in the gallery/probe set for each method. Thus, the result on the dataset from the FERET database is more convincing to evaluate the robustness of each method. From the above experimental results, it is very obvious that NDP is more stable than and superior to the other methods.

Table 1
Comparison of top recognition rate and corresponding dimension ℓ on the ORL database

Method	3 Train	4 Train	5 Train
PCA	88.55% (119)	92.12% (155)	94.40% (196)
LDA	86.78% (39)	90.67% (39)	92.17% (39)
LDE	87.28% (42)	90.77% (54)	92.85% (73)
SNPE	87.04% (40)	90.15% (40)	91.86% (40)
SLF	87.69% (39)	90.52% (39)	91.99% (39)
Direct-LDA	87.45% (39)	90.92% (39)	92.40% (39)
DCV	89.73% (39)	93.03% (39)	94.87% (39)
MFA	87.93% (40)	91.42% (48)	93.66% (59)
NDP	91.11% (36)	94.45% (39)	96.31% (39)

Furthermore, from the above three experiments, several aspects are worthwhile to emphasize:

(1) Different from PCA, LDA, Direct-LDA and DCV which attempt to preserve the *global* Euclidean structure, NDP aims at preserving the within-class *neighboring* geometry, which is more important than the global Euclidean structure in many practical classification problems (He et al., 2005a). Furthermore, NDP makes use of the between-class neighboring information to improve the discrimination ability of projected vectors.

(2) In contrast to SNPE and SLF which aim at preserving the local manifold structure, NDP aims at extracting a linear subspace, which not only preserves the within-class neighboring geometry but also makes the projected vectors of the samples of different classes far from each other. Therefore, NDP is better than SNPE and SLF for classification problems.

(3) Although MFA and LDE consider the within-class information and between-class information to address classification problems, they use PCA for dimensionality reduction to deal with the small sample size problem, which may result in the loss of important discriminative information. Different from MFA and LDE, NDP uses the null space of the within-class neighborhood scatter matrix to extract the linear facial subspace. Since the null space

Fig. 4. Twenty facial images of one individual in the UMIST database.

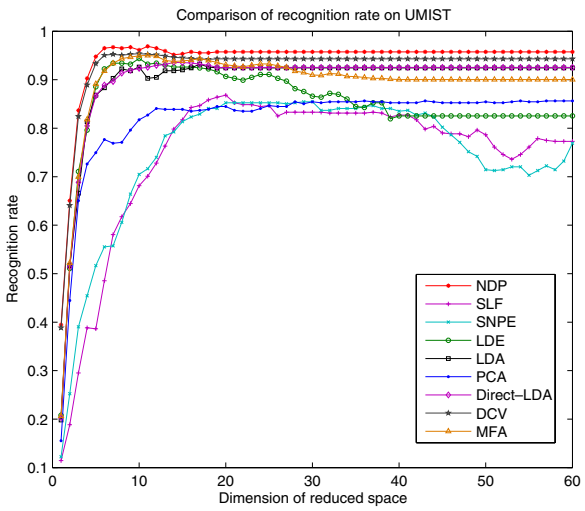

Fig. 5. Recognition rate versus dimension ℓ of reduced space on the UMIST database.

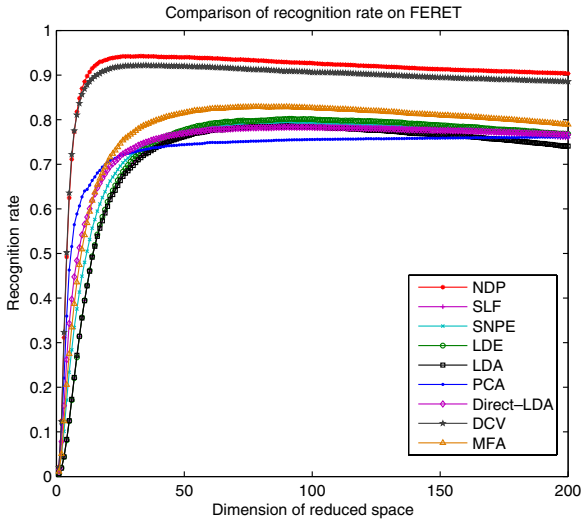

Fig. 7. Recognition rate versus dimension ℓ of reduced space on the FERET database.

Table 2
Comparison of top recognition rate and corresponding dimension ℓ on the UMIST database

Method	Dims	Top recognition rate (%)
PCA	36	85.63
LDA	17	93.20
LDE	10	94.37
SNPE	29	85.44
SLF	20	86.80
Direct-LDA	14	93.58
DCV	10	95.37
MFA	11	95.02
NDP	11	96.89

Table 3
Comparison of top recognition rate and corresponding dimension ℓ on the FERET database

Method	Dims	Top recognition rate (%)
PCA	436	75.14
LDA	90	78.58
LDE	92	80.21
SNPE	93	79.27
SLF	91	79.21
Direct-LDA	89	78.34
DCV	35	92.21
MFA	80	82.88
NDP	33	94.25

Fig. 6. Twenty facial images of 10 individuals in the FERET database.

method can extract more discriminant information for the small sample size problem, NDP is superior to MFA and LDE for face recognition.

(4) From the above experimental result, we can find that DCV is inferior to NDP but superior to the other methods. Both DCV and NDP use the null space to

extract the facial subspace; however, NDP uses the within-class neighborhood scatter matrix to compute the null space instead of using the within-class scatter matrix. Therefore, NDP can preserve the within-class neighboring geometry, and then obtains the better performance than DCV.

(5) NDP is *linear* and defined on *all* the training and testing samples, which makes it suitable for practical classification problems. Moreover, similar to KPCA (Schölkopf et al., 1998) generalized from PCA, we can generalize NDP to kernel NDP.

4. Conclusions

In this paper, we propose a novel manifold learning method named Neighborhood Discriminant Projection for face recognition. In order to preserve the within-class neighboring geometry of the image space and make the projected vectors of the samples of different classes far from each other, NDP explicitly considers the within-class submanifold and the between-class submanifold by the within-class graph and the between-class graph. Experimental results on ORL database, UMIST database and FERET database show the effectiveness and robustness of our proposed method. In future study, we will generalize the NDP to kernel NDP and further study the characteristic of the null space method.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 60635050 and 60405004, the National High-Tech Research and Development Plan of China under Grant Nos. 2006AA01Z318, 2006AA01Z192 and 20060101Z1059, and the National Basic Research Program of China under Grant No. 2006CB708303.

References

Athitsos, V., Alon, J., Sclaroff, S., Kollios, G., 2004. BoostMap: A method for efficient approximate similarity rankings. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 268–275.

Belhumeur, P.N., Hepanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE. Trans. Pattern Anal. Machine Intell. 19 (7), 711–720.

Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proc. Conf. on Advances in Neural Information Processing System, 14.

Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A., 2005. Discriminative common vectors for face recognition. IEEE. Trans. Pattern Anal. Machine Intell. 27 (1), 4–13.

Chellappa, R., Wilson, C.L., Sirohey, S., 1995. Human and machine recognition of faces: A survey. Proc. IEEE 83 (5), 705–741.

Chen, H.-T., Chang, H.-W., Liu, T.-L., Local discriminant embedding and its variants. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 846–853.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, second ed.. In: Computer Science and Scientific Computing Series Academic Press.

Graham, B.D., Allinson, M.N., 1998. Characterizing virtual eigensignatures for general purpose face recognition. In: Wechsler, H., Phillips, P.J., Bruce, V., Fogelman-Soulie, F., Huang, T.S. (Eds.), Face Recognition: From Theory to Applications. Springer-Verlag, Berlin, pp. 446–456.

He, X., Niyogi, P., 2003. Locality preserving projections. In: Proc. Conf. Advances in Neural Information Processing System 16.

He, X., Yan, S., Hu, Y., Zhang, H.-J., 2003. Learning a locality preserving subspace for visual recognition. In: Proc. Ninth IEEE Internat. Conf. on Computer Vision, vol. 1, pp. 385–392.

He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.-J., 2005a. Face recognition using Laplacianfaces. IEEE. Trans. Pattern Anal. Machine Intell. 27 (3), 328–340.

He, X., Cai, D., Yan, S., Zhang, H.-J., 2005b. Neighborhood preserving embedding. In: Proc. Tenth IEEE Internat. Conf. on Computer Vision, vol. 2, pp. 1208–1213.

Huang, R., Liu, O., Lu, H., Ma, S., 2002. Solving the small size problem of LDA. In: Proc. 16th Internat. Conf. Pattern Recognition, vol. 3, pp. 29–32.

Hu, C., Chang, Y., Feris, R., Turk, M., 2004. Manifold based analysis of facial expression. In: Proc. Computer Vision and Pattern Recognition Workshop, p. 81.

Li, S.Z., Jain, A.K., 2005. Handbook of Face Recognition, first ed. Springer.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J., 2000. The FERET evaluation methodology for face-recognition algorithms. IEEE. Trans. Pattern Anal. Machine Intell. 22 (10), 1090–1104.

Roweis, S., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Samaria, F.S., Harter, A.C., 1994. Parameterisation of a stochastic model for human face identification. In: Proc. 2nd IEEE Workshop on Applications of Computer Vision, pp. 138–142.

Saul, L.K., Roweis, S., 2003. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. J. Machine Learning Res. 4, 119–155.

Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10, 1299–1319.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Turk, M., 2001. A random walk through eigenspace. IEICE Trans. Inform. Syst. E84-D (12), 1586–1595.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. J. Cognitive Neurosci. 3 (1), 71–86.

Yan, S., Zhang, H., Hu, Y., Zhang, B., Cheng, Q., 2004. Discriminant analysis on embedded manifold. In: Proc. 8th European Conf. on Computer Vision, vol. 1, pp. 121–132.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., 2005. Graph embedding: A general framework for dimensionality reduction. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 20–25.

Yang, M.-H., 2002a. Kernel eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods. In: Proc. Fifth IEEE Internat. Conf. on Automatic Face and Gesture Recognition.

Yang, M.H., 2002b. Face recognition using extended isomap. In: Proc. Conf. Image Processing, vol. 2, pp. 117–120.

Yu, H., Yang, J., 2001. A direct LDA algorithm for high-dimensional data with application to face recognition. Pattern Recognition 34, 2067–2070.

Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A., 2003. Face recognition: A literature survey. ACM Comput. Surveys 35 (4), 399–459.