Subspace Semi-supervised Fisher Discriminant Analysis

YANG Wu-Yi[1, 2, 3] LIANG Wei[1] XIN Le[4] ZHANG Shu-Wu[1]

Abstract Fisher discriminant analysis (FDA) is a popular method for supervised dimensionality reduction. FDA seeks for an embedding transformation such that the ratio of the between-class scatter to the within-class scatter is maximized. Labeled data, however, often consume much time and are expensive to obtain, as they require the efforts of human annotators. In order to cope with the problem of effectively combining unlabeled data with labeled data to find the embedding transformation, we propose a novel method, called subspace semi-supervised Fisher discriminant analysis (SSFDA), for semi-supervised dimensionality reduction. SSFDA aims to find an embedding transformation that respects the discriminant structure inferred from the labeled data and the intrinsic geometrical structure inferred from both the labeled and unlabeled data. We also show that SSFDA can be extended to nonlinear dimensionality reduction scenarios by applying the kernel trick. The experimental results on face recognition demonstrate the effectiveness of our proposed algorithm.

Key words Fisher discriminant analysis (FDA), semi-supervised learning, manifold regularization, dimensionality reduction

In cases of machine learning and data mining, such as image retrieval, and face recognition, we may increasingly confront with the collection of high-dimensional data. This leads us to consider methods of dimensionality reduction that allow us to represent the data in a lower dimensional space. Techniques for dimensionality reduction have attracted much attention in computer vision and pattern recognition. The most popular dimensionality reduction algorithms include principal component analysis (PCA)[1−2] and Fisher discriminant analysis (FDA)[3].

PCA is an unsupervised method. It projects the original m-dimensional data into a d ($d \ll m$)-dimensional subspace in which the data variance is maximized. It computes the eigenvectors of the data covariance matrix, and approximates the original data by a linear combination of the leading eigenvectors. If the data are embedded in a linear subspace, PCA is guaranteed to discover the dimensionality of the subspace and produces a compact representation.

Unlike PCA, FDA is a supervised method. In the context of pattern classification, FDA seeks for the best projection subspace such that the ratio of the between-class scatter to the within-class scatter is maximized. For classification task, FDA can achieve significant better performance than PCA.

Labeled data, however, often consume much time and are expensive to obtain, as they require the efforts of human annotators[4]. Contrarily, in many cases, it is far easier to obtain large numbers of unlabeled data. The problem of effectively combining unlabeled data with labeled data is therefore of central importance in machine learning[4]. Learning from labeled and unlabeled data has attracted an increasing amount of attention recently, and several novel approaches have been proposed. Graph-based semi-supervised learning algorithms[4−13] have attracted considerable attention in recent years. These algorithms consider the graph over all the data as a priori knowledge to guide the decision making. The regularization-based technique of Cai[8] is closest in spirit to the intuitions of our paper. Techniques of Belkin[5] and Cai[8] are based on regularization.

In this paper, we aim at dimensionality reduction in semi-supervised case. To cope with the problem of effectively combining unlabeled data with labeled data, we propose a novel semi-supervised dimensionality reduction algorithm called subspace semi-supervised Fisher discriminant analysis (SSFDA). SSFDA exploits the geometric structure of the labeled an unlabeled data and incorporates it as an additional regularization term. SSFDA intends to find an embedding transformation that respects the discriminant structure inferred from the labeled data and the intrinsic geometrical structure inferred from both labeled and unlabeled data.

Semi-supervised discriminant analysis (SDA)[8] is the most relevant algorithm to our algorithm. In the following, we list the similarities and major difference between SDA and our algorithm:

1) Both SDA and our algorithm are graph-based approaches. Both use a p-nearest neighbor graph to model the relationship between the nearby data points and incorporate the geometric structure of the labeled and unlabeled data as an additional regularization term.

2) There is one major difference between SDA and our algorithm. In the SDA algorithm, without considering the labels of the labeled data, the weight matrix of the p-nearest neighbor graph is constructed according to the relationship between nearby points in the original data space. In our algorithm, using the labeled data, we first find a projection subspace by applying the FDA algorithm and embed the labeled and unlabeled data into this subspace. Then, the weight matrix of the p-nearest neighbor graph is constructed according to the relationship between nearby data points in the subspace, as well as the labels of the labeled data.

The rest of this paper is organized as follows. In Section 1, we provide a brief review of FDA. The proposed SSFDA algorithm for dimensionality reduction is introduced in Section 2. The experimental results are presented in Section 3. Finally, we conclude the paper in Section 4.

1 Fisher discriminant analysis

In this section, we first formulate the problem of linear dimensionality reduction. Then, FDA is reviewed. Last, the graph perspective of FDA is introduced.

Received March 18, 2008; in revised form June 8, 2009
Supported by National Sciences and Technology Supporting Program of China (2008BAH26B02-3, 2008BAH21B03-04, 2008BAH26B03)

1. Hi-tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China 2. Key Laboratory of Underwater Acoustic Communication and Marine Information Technology of the Minister of Education, Xiamen University, Xiamen 361005, P. R. China 3. College of Oceanography and Environmental Science, Xiamen University, Xiamen 361005, P. R. China 4. School of Electronics Information and Control Engineering, Beijing University of Technology, Beijing 100124, P. R. China

DOI: 10.3724/SP.J.1004.2009.01513

1.1 Formulation

Suppose that we have a set of l samples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_l \in \mathbf{R}^m$ that belong to c classes, then $l = \sum_{k=1}^{c} l_k$, where l_k is the number of samples in the k-th class. For the linear dimensionality reduction, we focus on finding a transformation matrix $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_d)$ that maps these l points to a set of points $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_l$ in $\mathbf{R}^d$ $(d \ll m)$. The embedded sample $\boldsymbol{y}_i$ is given by $\boldsymbol{y}_i = A^{\mathrm{T}}\boldsymbol{x}_i$.

1.2 Fisher discriminant analysis for dimensionality reduction

FDA[3] is one of the most popular dimensionality reduction techniques. FDA seeks directions on which the data points of different classes are far from each other while data points of the same class are close to each other[3]. Here, we briefly describe the definition of FDA.

Let S_w, S_b, and S_t be the within-class scatter matrix, the between-class scatter matrix, and total scatter matrix, respectively.

$$S_w = \sum_{k=1}^{c}\sum_{i=1}^{l_k} (\boldsymbol{x}_i^{(k)} - \boldsymbol{\mu}^{(k)})(\boldsymbol{x}_i^{(k)} - \boldsymbol{\mu}^{(k)})^{\mathrm{T}} \tag{1}$$

$$S_b = \sum_{k=1}^{c} l_k(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu})^{\mathrm{T}} \tag{2}$$

$$S_t = \sum_{k=1}^{c}\sum_{i=1}^{l_k} (\boldsymbol{x}_i^{(k)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(k)} - \boldsymbol{\mu})^{\mathrm{T}} = S_w + S_b \tag{3}$$

where l_k is the number of data in the k-th class, $\boldsymbol{x}_i^{(k)}$ is the i-th data in the k-th class, $\boldsymbol{\mu}^{(k)}$ is the mean of the data in the k-th class, and $\boldsymbol{\mu}$ is the mean of all data:

$$\boldsymbol{\mu}^{(k)} = \frac{1}{l_k}\sum_{i=1}^{l_k} \boldsymbol{x}_i^{(k)} \tag{4}$$

$$\boldsymbol{\mu} = \frac{1}{l}\sum_{i=1}^{l} \boldsymbol{x}_i \tag{5}$$

The objective function of FDA is as follows:

$$\boldsymbol{a}_{\mathrm{opt}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^{\mathrm{T}}S_b\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}S_w\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^{\mathrm{T}}S_b\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}S_t\boldsymbol{a}} \tag{6}$$

The projection vector $\boldsymbol{a}$ that maximizes (6) is given by the maximum eigenvalue solution to the generalized eigenvalue problem:

$$S_b\boldsymbol{a} = \lambda S_t\boldsymbol{a} \tag{7}$$

Let the column vector $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_d$ be the solutions of (7), ordered according to their eigenvalues, $\lambda_1 > \cdots > \lambda_d$. Thus, the embedding is as follows:

$$\boldsymbol{x}_i \to \boldsymbol{y}_i = A^{\mathrm{T}}\boldsymbol{x}_i$$

where $\boldsymbol{y}_i$ is a d-dimensional representation of the high dimensional data point $\boldsymbol{x}_i$ and $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_d)$ is the transformation matrix. The between-class scatter matrix S_b has at most rank $c - 1$. This implies that the multiplicity of $\lambda = 0$ is at least $m - c + 1$. Therefore, FDA can find at most $c - 1$ meaningful directions.

1.3 Graph perspective of Fisher discriminant analysis

We have

$$\begin{aligned} S_b = \sum_{k=1}^{c} l_k(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu})^{\mathrm{T}} = \\ \sum_{k=1}^{c} l_k \left(\frac{1}{l_k}\sum_{i=1}^{l_k} \boldsymbol{x}_i^{(k)}\right)\left(\frac{1}{l_k}\sum_{i=1}^{l_k} \boldsymbol{x}_i^{(k)}\right)^{\mathrm{T}} - l\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} = \\ \sum_{k=1}^{c} X^{(k)}W^{(k)}(X^{(k)})^{\mathrm{T}} - l\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \end{aligned} \tag{8}$$

$$S_t = \sum_{k=1}^{c}\sum_{i=1}^{l_k} (\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^{\mathrm{T}} - l\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \tag{9}$$

where $W^{(k)}$ is an $l_k \times l_k$ matrix with all the elements equal to $1/l_k$ and $X^{(k)} = [\boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_{l_k}^{(k)}]$. Let $X_l = [X^{(1)}, \cdots, X^{(c)}]$ and define an $l \times l$ matrix $W_{l\times l}$ as

$$W_{l\times l} = \begin{bmatrix} W^{(1)} & 0 & \cdots & 0 \\ 0 & W^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W^{(c)} \end{bmatrix} \tag{10}$$

Then, we have

$$S_b = X_lBX_l^{\mathrm{T}} \tag{11}$$

$$S_t = X_lCX_l^{\mathrm{T}} \tag{12}$$

$$B = W_{l\times l} - \frac{1}{l}\mathbf{1}\mathbf{1}^{\mathrm{T}}$$

$$C = I - \frac{1}{l}\mathbf{1}\mathbf{1}^{\mathrm{T}}$$

where $\mathbf{1} = [1, \cdots, 1]^{\mathrm{T}}$ is an l-dimensional vector and I is an $l \times l$ identity matrix. Thus, the objective function of FDA in (6) can be rewritten as

$$\boldsymbol{a}_{\mathrm{opt}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^{\mathrm{T}}S_b\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}S_t\boldsymbol{a}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^{\mathrm{T}}X_lBX_l^{\mathrm{T}}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}X_lCX_l^{\mathrm{T}}\boldsymbol{a}} \tag{13}$$

The generalized eigenvalue problem in (7) can be rewritten as

$$X_lBX_l^{\mathrm{T}}\boldsymbol{a} = \lambda X_lCX_l^{\mathrm{T}}\boldsymbol{a} \tag{14}$$

This formulation of FDA objective function was first introduced in [14].

2 Subspace semi-supervised Fisher discriminant analysis

We introduce our SSFDA algorithm that respects both discriminant and geometrical structures in the data. We begin with a description of the semi-supervised learning problem.

2.1 Problem formulation

Given a sample set $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \cdots, \boldsymbol{x}_n\} \subset \mathbf{R}^m$ and a label set $L = \{1, \cdots, c\}$, the first l points $\boldsymbol{x}_i$ $(i \le l)$ are labeled as $t_i \in L$ and the remaining points $\boldsymbol{x}_u (l + 1 \le u \le n)$ are unlabeled. Find a transformation matrix $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_d)$ that maps these n points $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ to a set of points in $\mathbf{R}^d$ $(d \ll m)$. The embedded sample $\boldsymbol{y}_i$ is given by $\boldsymbol{y}_i = A^{\mathrm{T}}\boldsymbol{x}_i$. For any unlabeled sample

$\boldsymbol{x}_u$ $(l+1 \leq u \leq n)$, its label is then predicted as t_{i^*} provided that $\boldsymbol{y}_{i^*} = A^{\mathrm{T}}\boldsymbol{x}_{i^*}$ minimizes $\|\boldsymbol{y}_i - \boldsymbol{y}_u\|$, $i = 1, \cdots, l$. The performance of an algorithm is measured by the recognition error rate on the unlabeled samples.

2.2 The objective function

FDA is a supervised method. It wants to find an embedding transformation such that the ratio of the between-class scatter to the within-class scatter is maximized. When there is no sufficient training (labeled) samples, the problem of learning from both labeled and unlabeled samples (semi-supervised learning) is of central importance in improving the performance of recognition.

SSFDA will be performed in the subspace $\mathbf{R}(X_k)$, where $\mathbf{R}(X_k) = \mathrm{Span}[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k]$ is the subspace spanned by the columns of $X_k = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k]$ and $k = l$ or n. Suppose that the rank of X_k is r, i.e., $\mathrm{rank}(X_k) = \dim(\mathbf{R}(X_k)) = r$. Perform the singular value decomposition of X_k as $X_k = U\sum V^{\mathrm{T}}$, where $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r, \boldsymbol{u}_{r+1}, \cdots, \boldsymbol{u}_m]$ is the left orthonormal matrix, $\sum$ is the singular value matrix, and V is the right orthonormal matrix. The subspace $\mathbf{R}(X_k)$ can be spanned by the columns of P, where $P = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$. When $\mathrm{rank}(X_k) = m$, we can simply select the $m \times m$ identity matrix I as P, i.e., $P = I$. We project both labeled and unlabeled data into subspace $\mathbf{R}(X_k)$. Thus, the embedding is as follows:

$$\boldsymbol{x}_i \rightarrow \boldsymbol{z}_i = P^{\mathrm{T}}\boldsymbol{x}_i, \quad i = 1, \cdots, l, l+1, \cdots, n$$

Let $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$, $X_l = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_l]$, $Z^{(k)} = [\boldsymbol{z}_1^{(k)}, \cdots, \boldsymbol{z}_{l_k}^{(k)}]$, $Z_l = [Z^{(1)}, \cdots, Z^{(c)}]$, and $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots \boldsymbol{z}_n]$. Then, $Z_l = PX_l$ and $Z = PX$. In the subspace, the between-class scatter matrix S_b and total scatter matrix S_t are as follows:

$$S_b = Z_l B Z_l^{\mathrm{T}} = PX_l B X_l^{\mathrm{T}} P^{\mathrm{T}} \tag{15}$$

$$S_t = Z_l C Z_l^{\mathrm{T}} = PX_l C X_l^{\mathrm{T}} P^{\mathrm{T}} \tag{16}$$

Then, the objective function of FDA is as follows:

$$\boldsymbol{a}_{\mathrm{opt}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^{\mathrm{T}} PX_l B X_l^{\mathrm{T}} P^{\mathrm{T}} \boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}} PX_l C X_l^{\mathrm{T}} P^{\mathrm{T}} \boldsymbol{a}}$$

The optimal $\boldsymbol{a}$ is the eigenvector corresponding to the maximum eigenvalue of eigen-problem:

$$Z_l B Z_l^{\mathrm{T}} \boldsymbol{a} = \lambda Z_l C Z_l^{\mathrm{T}} \boldsymbol{a} \tag{17}$$

A typical way to prevent overfitting, which may happen when there is no sufficient training sample, is to impose a regularizer[15]. The optimization problem of the regularized version of FDA can be written as follows:

$$\boldsymbol{a}_{\mathrm{opt}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^{\mathrm{T}} S_b \boldsymbol{a}}{\alpha \cdot \boldsymbol{a}^{\mathrm{T}} S_t \boldsymbol{a} + (1-\alpha) R(\boldsymbol{a})} \tag{18}$$

The a priori knowledge of the data can be incorporated in the regularized term $R(\boldsymbol{a})$, which provides us the flexibility to incorporate the geometrical structure of the data manifold[8]. The key to semi-supervised learning problems is the priori assumption of consistency[7], which means that 1) nearby points are likely to have the same label and 2) points on the same structure (typically referred to as a cluster or a manifold) are likely to have the same label. A principled approach to formalize the assumption of consistency is to design a mapping function which is sufficiently smooth with respect to the intrinsic structure revealed by known labeled and unlabeled points[7]. In order to discover both geometrical and discriminant structures of the data manifold, we incorporate the manifold structure of both labeled and unlabeled data as the regularization term in the objective function (18).

When $k = n$, S_t in (16) may be singular. We use regularization to ensure the nonsingularity of S_t: $S_t = S_t + \delta I_r$, where δ $(\delta > 0)$ is the regularization parameter and I_r is the $r \times r$ identity matrix. Let the column vectors $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{c-1}$ be the solutions of (17), ordered according to their eigenvalues, $\lambda_1 > \cdots > \lambda_{c-1}$. Then, we find a transformation matrix $B = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{c-1}]$. The $r \times (c-1)$ transformation matrix B maps all the n samples to a set of points $\{\boldsymbol{q}_i | \boldsymbol{q}_i = B^{\mathrm{T}}\boldsymbol{z}_i, i = 1, \cdots, n\}$ in $\mathbf{R}^{c-1}$. Let $l(\boldsymbol{x}_i)$ be the class label of $\boldsymbol{x}_i$, and $N_p(\boldsymbol{q}_i) = \{\boldsymbol{q}_i^1, \boldsymbol{q}_i^2, \cdots, \boldsymbol{q}_i^p\}$ be the set of $\boldsymbol{q}_i'$s p-nearest neighbors. We construct a p-nearest neighbor graph G to model the relationship between nearby data points. Thus, the weight matrix W of G can be defined as follows:

$$W_{ij} = \begin{cases} \gamma, \text{ if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are labeled, and} \\ \quad \text{they share the same label;} \\ \gamma, \text{ if } \boldsymbol{x}_i \text{ is labeled, } \boldsymbol{x}_j \text{ is unlabeled,} \\ \quad \boldsymbol{q}_i \in N_p(\boldsymbol{q}_j), \text{ and for each } \boldsymbol{q}_s \in N_p(\boldsymbol{q}_j), \\ \quad \text{if } \boldsymbol{x}_s \text{ is labeled, } l(\boldsymbol{x}_s) = l(\boldsymbol{x}_i); \\ \gamma, \text{ if } \boldsymbol{x}_j \text{ is labeled, } \boldsymbol{x}_i \text{ is unlabeled,} \\ \quad \boldsymbol{q}_j \in N_p(\boldsymbol{q}_i) \text{ and for each } \boldsymbol{q}_s \in N_p(\boldsymbol{q}_i), \\ \quad \text{if } \boldsymbol{x}_s \text{ is labeled, } l(\boldsymbol{x}_s) = l(\boldsymbol{x}_j); \\ 1, \text{ if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are unlabeled,} \\ \quad \text{and } \boldsymbol{q}_i \in N_p(\boldsymbol{q}_j) \text{ or } \boldsymbol{q}_j \in N_p(\boldsymbol{q}_i); \\ 0, \text{ otherwise} \end{cases} \tag{19}$$

In general, if two data points are linked by an edge, they are likely to be in the same class. Thus, a natural regularizer can be defined as follows:

$$\begin{aligned} R(\boldsymbol{a}) = \frac{1}{2}\sum_{i,j} (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{z}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{z}_j)^2 W_{ij} = \\ \sum_i \boldsymbol{a}^{\mathrm{T}}\boldsymbol{z}_i D_{ii} \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{a} - 2\sum_{i,j} \boldsymbol{a}^{\mathrm{T}}\boldsymbol{z}_i W_{ij} \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{a} = \\ \boldsymbol{a}^{\mathrm{T}} Z(D-W)Z^{\mathrm{T}}\boldsymbol{a} = \boldsymbol{a}^{\mathrm{T}} PXLX^{\mathrm{T}}P^{\mathrm{T}}\boldsymbol{a} \end{aligned} \tag{20}$$

where D is a diagonal matrix; its entries are column sum of W, $D_{ii} = \sum_j W_{ij}$, and $L = D - W$ is the Laplacian matrix. We get the objective function of SSFDA:

$$\boldsymbol{a}_{\mathrm{opt}} = \arg\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^{\mathrm{T}} S_b \boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}(\alpha S_t + (1-\alpha)PXLX^{\mathrm{T}}P^{\mathrm{T}})\boldsymbol{a}} \tag{21}$$

The projective vector $\boldsymbol{a}$ that maximizes (21) is given by the maximum eigenvalue solution to the generalized eigenvalue problem:

$$S_b\boldsymbol{a} = \lambda(\alpha S_t + (1-\alpha)PXLX^{\mathrm{T}}P^{\mathrm{T}})\boldsymbol{a} \tag{22}$$

$$PX_l B X_l^{\mathrm{T}} P^{\mathrm{T}}\boldsymbol{a} = \lambda P(\alpha X_l C X_l^{\mathrm{T}} + (1-\alpha)XLX^{\mathrm{T}})P^{\mathrm{T}}\boldsymbol{a} \tag{23}$$

Let the column vectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_d$ be the solutions of (23), ordered according to their eigenvalues, $\lambda_1 > \cdots > \lambda_d$. Thus, the embedding is as follows:

$$\boldsymbol{x}_i \rightarrow \boldsymbol{y}_i = A^{\mathrm{T}}\boldsymbol{z}_i = A^{\mathrm{T}}P^{\mathrm{T}}\boldsymbol{x}_i = (PA)^{\mathrm{T}}\boldsymbol{x}_i$$

where $\boldsymbol{y}_i$ is a d-dimensional representation of the high dimensional data point $\boldsymbol{x}_i$ and $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_d)$. PA is the transformation matrix. The between-class scatter matrix S_b has at most rank $c-1$. This implies that the multiplicity of $\lambda = 0$ is at least $m-c+1$. Therefore, SSFDA can find at most $c-1$ meaningful directions.

2.3 Algorithm

In summary of the discussion so far, the SSFDA algorithm is given below:

1) Find a orthonormal basis of the subspace $\mathbf{R}(X_k)$: Perform singular value decomposition of X_k as $X_k = U\sum V^{\mathrm{T}}$. $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r, \boldsymbol{u}_{r+1}, \cdots, \boldsymbol{u}_m]$ is the orthonormal matrix, where r is the rank of X_k. The vector set $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r\}$ forms a orthonormal basis of $\mathbf{R}(X_k)$. The matrix $P = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$ gives a vector space projection from $\mathbf{R}^m$ to subspace $\mathbf{R}(X_k)$.

2) Map data points into subspace $\mathbf{R}(X_l)$:

$$\boldsymbol{x}_i \rightarrow \boldsymbol{z}_i = P^{\mathrm{T}}\boldsymbol{x}_i, \quad i = 1, \cdots, l, l+1, \cdots, n$$

3) Construct the scatter matrixes: Construct the between-class scatter matrix S_b and total scatter matrix S_t as

$$S_b = Z_l B Z_l^{\mathrm{T}} = P X_l B X_l^{\mathrm{T}} P^{\mathrm{T}}$$

$$S_t = Z_l C Z_l^{\mathrm{T}} = P^{\mathrm{T}} X_l C X_l^{\mathrm{T}} P^{\mathrm{T}}$$

4) Eigen-problem: When $k = n$, we use regularization to ensure the nonsingularity of S_t: $S_t = S_t + \delta I_r$, where $\delta\,(\delta > 0)$ is the regularization parameter and I_r is the $r \times r$ identity matrix. Compute the eigenvectors with respect to the nonzero eigenvalues for the generalized eigenvector problem:

$$S_b\boldsymbol{b}_i = \lambda_i S_t \boldsymbol{b}_i, \quad i = 1, 2, \cdots, c-1$$

We obtain the transformation matrix $B = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{c-1}]$.

5) Construct the adjacency graph: Construct the p-nearest neighbor graph matrix W as in (19) to model the relationship between nearby data points and calculate the graph Laplacian matrix $L = D - W$.

6) Eigen-problem: Compute the eigenvectors with respect to the nonzero eigenvalues for the generalized eigenvector problem:

$$S_b\boldsymbol{a}_i = \lambda(\alpha S_t + (1-\alpha)PXLX^{\mathrm{T}}P^{\mathrm{T}})\boldsymbol{a}_i, \quad i = 1, \cdots, c-1$$

There are at most $c-1$ nonzero eigenvalues for the eigenvector problem. We obtain the transformation matrix $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{c-1})$.

7) SSFDA embedding: The data point can be embedded into $c-1$ dimensional subspace by

$$\boldsymbol{x} \rightarrow \boldsymbol{y} = A^{\mathrm{T}}\boldsymbol{z} = A^{\mathrm{T}}P^{\mathrm{T}}\boldsymbol{x} = (PA)^{\mathrm{T}}\boldsymbol{x}$$

2.4 Kernel SSFDA for nonlinear dimensionality reduction

When the data manifold is highly nonlinear, the linear method described above may fail to discover the intrinsic geometry. Here, we show how SSFDA can be extended to nonlinear dimensionality reduction scenarios.

For a given nonlinear mapping ϕ, the input data in the space $\mathbf{R}^m$ can be mapped into the space $\mathcal{H}$:

$$\phi : \mathbf{R}^m \rightarrow \mathcal{H}, \boldsymbol{x} \mapsto \phi(\boldsymbol{x})$$

Kernel SSFDA will be performed in the subspace $\mathbf{R}(\Phi_k)$, where $\Phi_k = [\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_k)]$ and $k = l$ or n. Let $\{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_r\}$ be an orthonormal basis of the subspace $\mathbf{R}(\Phi_k)$, where r is the rank of Φ_k. It is easy to show that every vector $\boldsymbol{\beta}_i, i = 1, \cdots, r$ can be linearly expanded by

$$\boldsymbol{\beta}_i = \sum_{j=1}^{k} \gamma_{ij}\phi(\boldsymbol{x}_j), \quad i = 1, \cdots, r$$

Denote $M = \Phi_k^{\mathrm{T}}\Phi_k$, whose elements can be determined by the following kernel function:

$$M_{i,j} = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) = (\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j)) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Calculate the orthonormal eigenvectors $\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_r$ of M corresponding to the r largest positive eigenvalues, $\lambda_1 > \cdots > \lambda_r$. Then, the orthonormal vectors $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_r$ can be obtain by

$$\boldsymbol{\beta}_i = \lambda_i^{-\frac{1}{2}}\Phi_k\boldsymbol{\nu}_i, \quad i = 1, \cdots, r$$

After projection of the mapped samples $\phi(\boldsymbol{x}_i)$, $i = 1, \cdots, n$ into the subspace $\mathbf{R}(\Phi_k)$, the transformed vectors $\boldsymbol{z}_i$, $i = 1, \cdots, n$ can be obtained by

$$\boldsymbol{z}_i = P^{\mathrm{T}}\phi(\boldsymbol{x}_i), \quad i = 1, \cdots, l, l+1, \cdots, n \tag{24}$$

where $P = [\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_r]$. Let $\Lambda = \mathrm{diag}\{\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_r^{-\frac{1}{2}}\}$ and $V = [\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_r]$. Equation (24) can be rewritten as

$$\begin{aligned}\boldsymbol{z}_i = [\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_r]^{\mathrm{T}}\phi(\boldsymbol{x}_i) = \Lambda V^{\mathrm{T}}\Phi_k^{\mathrm{T}}\phi(\boldsymbol{x}_i) = \\ \Lambda V^{\mathrm{T}}[K(\boldsymbol{x}_1, \boldsymbol{x}_i), \cdots, K(\boldsymbol{x}_k, \boldsymbol{x}_i)]^{\mathrm{T}}\end{aligned} \tag{25}$$

Construct the between-class scatter matrix S_b and total scatter matrix S_t as

$$S_b = Z_l B Z_l^{\mathrm{T}}$$

$$S_t = Z_l C Z_l^{\mathrm{T}}$$

When $k = n$, S_t may be singular. We use regularization to ensure the nonsingularity of S_t: $S_t = S_t + \delta I_r$, where $\delta\,(\delta > 0)$ is the regularization parameter and I_r is the $r \times r$ identity matrix. Compute the eigenvectors with respect to the nonzero eigenvalues for the eigenvector problem:

$$S_b\boldsymbol{b}_i = \lambda S_t\boldsymbol{b}_i, \quad i = 1, 2, \cdots, c-1$$

We obtain the transformation matrix $B = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{c-1})$. The transformation matrix B maps all the n samples to a set of points $\{\boldsymbol{q}_i | \boldsymbol{q}_i = B^{\mathrm{T}}\boldsymbol{z}_i, i = 1, \cdots, n\}$ in $\mathbf{R}^{c-1}$. Construct the p-nearest neighbor graph matrix W as in (19) to model the relationship between nearby data points, and calculate the graph Laplacian matrix $L = D - W$.

Compute the eigenvectors with respect to the nonzero eigenvalues for the generalized eigenvector problem:

$$S_b\boldsymbol{a}_i = \lambda(\alpha S_t + (1-\alpha)ZLZ^{\mathrm{T}})\boldsymbol{a}_i, \quad i = 1, \cdots, c-1$$

Let $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{c-1})$. For a data point $\boldsymbol{x}$, the embedded data point $\boldsymbol{y}$ can be obtained by

$$\begin{aligned}&\boldsymbol{y} = A^{\mathrm{T}}\boldsymbol{z} = \ A^{\mathrm{T}}P^{\mathrm{T}}\phi(\boldsymbol{x}) = \\ &A^{\mathrm{T}}[\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_r]^{\mathrm{T}}\phi(\boldsymbol{x}) = \ A^{\mathrm{T}}\Lambda V^{\mathrm{T}}\Phi_k^{\mathrm{T}}\phi(\boldsymbol{x}) = \\ &(\mathrm{V}\Lambda A)^{\mathrm{T}}[K(\boldsymbol{x}_1, \boldsymbol{x}), \cdots, K(\boldsymbol{x}_k, \boldsymbol{x})]^{\mathrm{T}}\end{aligned}$$

3 Experimental results

In this section, the performance of the proposed SSFDA for face recognition is investigated.

A face recognition task is handled as a multi-class classification problem. Each test image is mapped to a low-dimensional subspace via the embedding learned from training data, and then it is classified by the nearest neighbor criterion.

3.1 Datasets and compared algorithms

In our experiments, we use the CMU PIE (Pose, Illumination and Expression) databases[16] for face recognition to evaluate our proposed SSFAD algorithm. The CMU PIE face database contains more than 40 000 facial images of 68 people. The face images were captured under varying pose, illumination, and expression. From the dataset that contains five near frontal poses (C05, C07, C09, C27, and C29), we randomly select 20 persons and 80 images for each person in the experiments.

In all the experiments, preprocessing to locate the faces was applied. Original images were normalized (in scale and orientation) such that the two eyes were aligned at the same position. Then, the facial areas were cropped into the final images for matching. The size of each cropped image in all the experiments is 32 × 32 pixels, with 256 gray levels per pixel. Thus, each image is represented by a 1 024-dimensional vector in image space.

The image set is then partitioned into the gallery and probe set with different numbers. For ease of representation, Gm/Ll/Pn means m images per person are randomly selected for training and the remaining n images are for testing. Among these m images, l images are randomly selected and labeled which leaves other $(m - l)$ images unlabeled.

We compare the performance of SSFDA with Fisherface[17] (PCA followed by FDA), Laplacianface[9] (PCA followed by LPP), and PCA + SDA (PCA followed by SDA)[8]. In order to deal with the "small sample size" problem, for Fisherface, Laplacianface, and PCA + SDA, we kept 99.9 % information in the sense of reconstruction error in the PCA step. We name the SSFDA as SSFDAl when the SSFDA is performed in the subspace $\mathbf{R}(X_l)$ and the SSFDA as SSFDAn when the SSFDA is performed in the subspace $\mathbf{R}(X_n)$. When the weight matrix W in the SSFDAn is obtained only according to the relationship between the data in the subspace spanned by the principle components after the PCA step, the SSFDAn degenerates to the PCA + SDA.

3.2 Face recognition results

The recognition error rates of different algorithms on CMU PIE databases are shown in Table 1. In each case, the minimum error rate is boldfaced. For each Gm/Ll/Pn, we average the results over 20 random splits and report the mean as well as the standard deviation.

Dimensionality estimation is a crucial problem for most of the subspace learning-based face recognition methods. The performance usually varies with the number of dimensions. We show the best results obtained by these subspace learning algorithms. The numbers in parentheses in Table 1 are the corresponding feature dimensions with the best results after dimensionality reduction.

From the results in Table 1, we can make the following comments: 1) The SSFDAl achieved the best performance of face recognition among all the compared algorithms, and the SSFDAn achieved a little poorer performance than the SSFDAl; 2) From the standard deviation of error rates shown in the table, we found that our algorithm is more stable than other algorithms; 3) The out-of-sample extension is good for SSFDA; 4) Compared with the performance of Fisherface, the proposed algorithm greatly improved the performance of face recognition by using the unlabeled data; 5) Compared with the performance of Fisherface, the SDA algorithm impaired the performance of face recognition by using the unlabeled data.

3.3 Model selection for SSFDA

Model selection is a crucial problem in many learning problems. If the learning performance varies drastically with different choices of the parameters, we have to apply some model selection methods for estimating the generalization error.

In our SSFDA algorithm, there are three parameters: α, γ, and p. In the previous experiments, we empirically set them as $\alpha = 0.8$, $\gamma = 0.9$, and $p = 5$. Fig. 1 shows the performance of SSFDAl with respect to different values of these parameters on CMU PIE database. Only one parameter was varied at a time while the others were fixed. As can be seen from Figs. 1 (a) and (b), when α is less than 0.9, the algorithm is not sensitive to α. However, when we gradually increased α from 0.9 to 1, the performance deteriorated gradually. When α is equal to 1, the SSFDA algorithm degenerates to the FDA algorithm and we get the worst performance. As can be seen from Figs. 1 (c) and (d), in order to get a good performance, we can empirically set γ as near 1. For the value of p, in order to discover the local geometrical and discriminant structures of the data space, it is usually set to a small number.

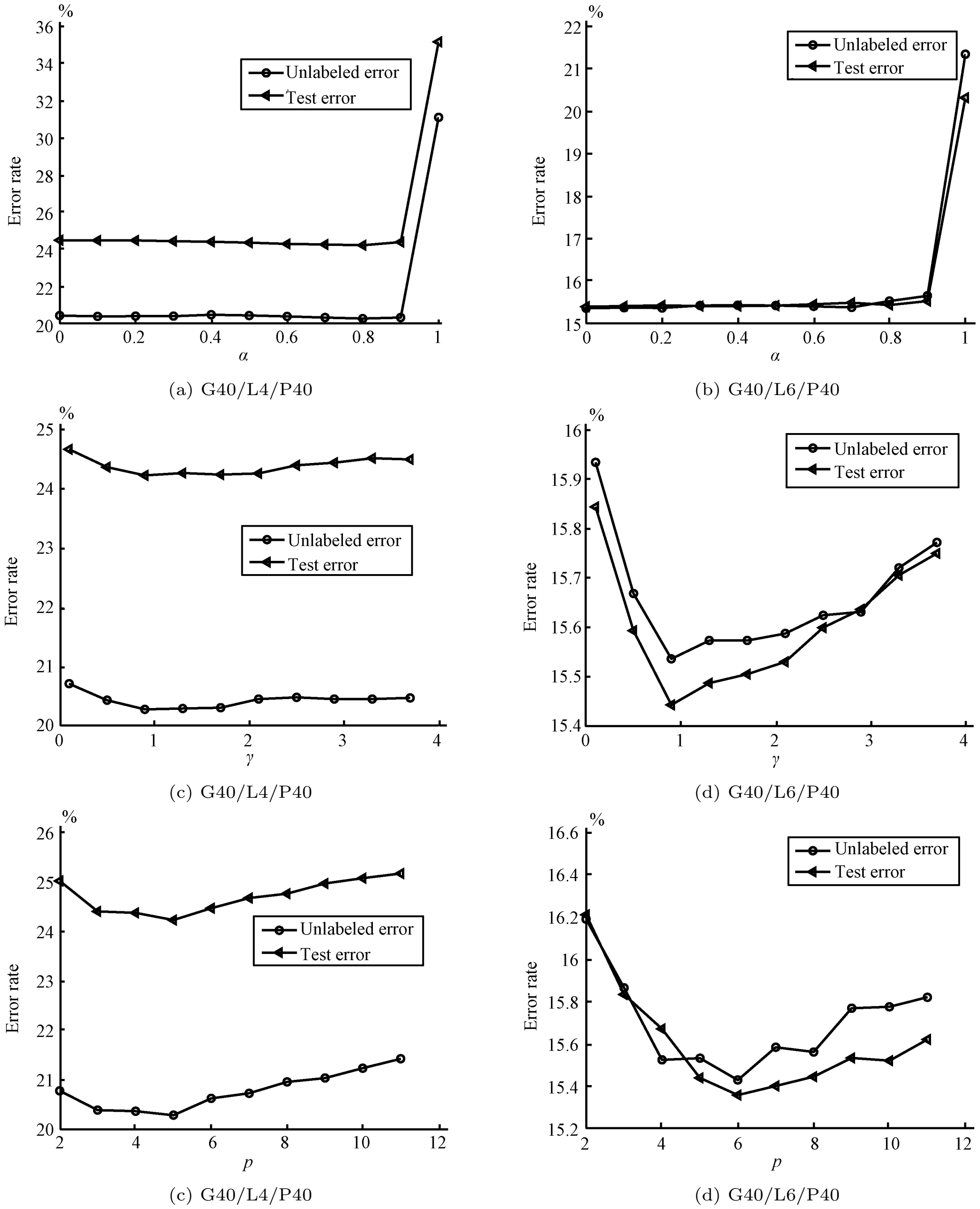
3.4 Discussion

Why could not the SDA algorithm improve the perfor-

Table 1 Recognition error rates on CMU PIE database (%)

Method	G40/L3/P40		G40/L4/P40	
	Unlabeled error	Test error	Unlabeled error	Test error
Laplacianface	38.0±8.3 (19)	38.4±10.3 (39)	30.4±10.7 (41)	35.2±9.2 (39)
Fisherface	40.6±8.8 (19)	41.6±10.3 (19)	31.4±10.2 (19)	36.1±8.6 (19)
PCA + SDA	45.5±7.0 (19)	43.7±7.5 (19)	38.6±10.5 (19)	40.5±8.5 (19)
SSFDAl	**26.2±6.3 (19)**	**26.3±7.5 (19)**	**20.2±8.0 (19)**	**24.2±6.0 (19)**
SSFDAn	28.8±7.2 (19)	29.5±9.3 (19)	22.7±9.4 (19)	26.0±7.5 (19)

Method	G40/L5/P40		G40/L6/P40	
	Unlabeled error	Test error	Unlabeled error	Test error
Laplacianface	27.1±10.3 (40)	27.5±10.7 (39)	22.1±6.7 (43)	21.1±5.8 (43)
Fisherface	28.5±10.6 (19)	28.8±11.2 (19)	23.1±7.6 (19)	22.2±6.2 (19)
PCA + SDA	37.3±10.0 (19)	36.6±10.7 (19)	33.9±7.8 (19)	30.4±7.4 (19)
SSFDAl	**19.0±7.7 (19)**	**19.7±8.7 (19)**	**15.5±6.0 (19)**	**15.4±4.3 (19)**
SSFDAn	20.0±7.9 (19	21.0±9.2 (19)	17.1±6.5 (19)	16.6±4.6 (19)



(a) G40/L4/P40

(b) G40/L6/P40

(c) G40/L4/P40

(d) G40/L6/P40

(c) G40/L4/P40

(d) G40/L6/P40

Fig. 1 Model selection for SSFDAl: error rates vs. different values of parameters α, γ, and p

mance of face recognition by using unlabeled data? Because the face images were captured under varying pose, illumination, and expression, the nearby data points in the original data space may belong to different classes. In the SDA algorithm, without considering the labels of the labeled data, the weight matrix of the p-nearest neighbor graph is constructed according to the relationship between nearby points in the original data space. The regularization term of SDA, which is based on the weight matrix, leads to seek for the projection subspace preserving the geometrical structure of the data. In the projection subspace found by the SDA algorithm, the mapped data of these nearby data, which belong to different classes, are still close to each other. So the SDA algorithm impaired the performance of face recognition instead of improving the performance.

In our algorithm, using the labeled data, we first find a projection subspace by applying the FDA algorithm and embed the labeled and unlabeled data into this subspace. In this subspace, the data points of different classes are apart from each other, while the data points of the same class are close to each other. So the nearby data in this subspace are more likely to belong to the same classes. Then, the weight matrix of the p-nearest neighbor graph is constructed according to the labels of the labeled data and the relationship between nearby points in the subspace. The regularization term of SSFDA respects the discriminant structure inferred from the labeled data and the intrinsic geometrical structure inferred from both labeled and unlabeled data. So, the proposed algorithm greatly improved the performance of face recognition by using the unlabeled data. We think the reason why the SSFDAl achieved a little better performance than the SSFDAn is that SSFDA can pay more attention to the discriminant structure in the subspace spanned by the labeled data.

4 Conclusion

In this paper, we propose a novel semi-supervised method for dimensionality reduction called subspace semi-supervised Fisher discriminant analysis. It can make efficient use of both labeled data and unlabeled data. SSFDA wants to find an embedding transformation that respects the discriminant structure inferred from the labeled data and the intrinsic geometrical structure inferred from both labeled and unlabeled data. Experimental results on face recognition have shown that the proposed algorithm is able to use unlabeled data effectively.

References

1 Mardia K V, Kent J T, Bibby J M. *Multivariate Analysis*. New York: Academic Press, 1980

2 Duda R O, Hart P E, Stork D G. *Pattern Classification (Second Edition)*. New Jersey: Wiley Interscience, 2000

3 Fukunaga K. *Introduction to Statistical Pattern Recognition (Second Edition)*. New York: Academic Press, 1990

4 Zhu X J, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 12th International Conference on Machine Learning. Washington D. C., USA: ACM, 2003. 912−919

5 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from examples. *Journal of Machine Learning Research*, 2006, **7**(11): 2399−2434

6 Sindhwani V, Niyogi P, Belkin M. Beyond the point cloud: from transductive to semi-supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany: ACM, 2005. 824−831

7 Zhou D, Bousquet O, Lal T N, Weston J, Scholkopf B. Learning with local and global consistency. *Advances in Neural Information Processing Systems 16*. Cambridge: MIT Press, 2003. 321−328

8 Cai D, He X F, Han J W. Semi-supervised discriminant analysis. In: Proceedings of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007. 1−7

9 Zha Z J, Mei T, Wang J D, Wang Z F, Hua X S. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 2009, **20**(2): 97−103

10 Zhao B, Wang F, Zhang C S, Song Y Q. Active model selection for graph-based semi-supervised learning. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: IEEE, 2008. 1881−1884

11 Johnson R, Zhang T. Graph-based semi-supervised learning and spectral kernel design. *IEEE Transactions on Information Theory*, 2008, **54**(1): 275−288

12 Gong Y C, Chen C L, Tian Y J. Graph-based semisupervised learning with redundant views. In: Proceedings of the 19th International Conference on Pattern Recognition. Tampa, USA: IEEE, 2008. 1−4

13 Luo J, Chen H, Tang Y. Analysis of graph-based semisupervised regression. In: Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery. Jinan, China: IEEE, 2008. 111−115

14 He X F, Yan S C, Hu Y X, Niyogi P, Zhang H J. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(3): 328−340

15 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001

16 Sim T, Baker S, Bsat M. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**(12): 1615−1618

17 Belhumeur P N, Hepanha J P, Kriegman D J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, **19**(7): 711−720

YANG Wu-Yi Assistant professor at the College of Oceanography and Environmental Science, Xiamen University. He received his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2009. His research interest covers acoustic signal processing, underwater acoustic data communication, image processing, video processing, character recognition, and pattern recognition. Corresponding author of this paper.
E-mail: eagleywy@126.com

LIANG Wei Assistant professor at the Institute of Automation, Chinese Academy of Sciences. He received his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2006. His research interest covers audio and video content analysis.
E-mail: wliang@hitic.ia.ac.cn

XIN Le Lecturer at the School of Electronics Information and Control Engineering, Beijing University of Technology. He received his M. S. degree from Xi′an Jiaotong University in 2003 and Ph. D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2008. His research interests covers intelligent information processing, computer vision, pattern recognition, intelligent transportation system, traffic image processing, and traffic simulation.
E-mail: xinle@bjut.edu.cn

ZHANG Shu-Wu Professor at the Institute of Automation, Chinese Academy of Sciences. He received his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 1997. He was an invited researcher in Advanced Telecomunications Research Institute Iternational, Spoken Language Translation Laboratories, Japan, from April in 1998 to May in 2002. His research interest covers digital media technologies, multilingual speech recognition, and natural language processing.
E-mail: swzhang@hitic.ia.ac.cn