Neurocomputing 91 (2012) 11–20

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

# Discriminative information preservation for face recognition

Dapeng Tao, Lianwen Jin *

*School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, PR China*

ARTICLE INFO

*Article history:*
Received 2 October 2011
Received in revised form
9 February 2012
Accepted 14 February 2012
Communicated by X. Gao
Available online 23 March 2012

*Keywords:*
Dimension reduction
Face recognition
Manifold learning
Patch alignment framework

ABSTRACT

It is usually difficult to find the optimal low dimensional subspace for face recognition. Patch alignment framework (PAF) is an important systematic framework that can be applied to understand the common thought and essential differences of a numerous dimensionality reduction algorithms, e.g., principal component analysis, linear discriminant analysis and locally linear embedding and ISOMAP. These algorithms do not consider the intra-class local geometry and the inter-class discrimination simultaneously. In this paper, we present a new dimensionality reduction algorithm based on PAF, termed the discriminative information preservation based dimensionality reduction or DIP for short. First, DIP models the local geometry of intra-class samples by using Locality preserving projection (LPP) rebuilt upon PAF. Second, it models the discriminative information of inter-class samples by maximizing the margin. Thoroughly experimental evidence on several public face datasets suggests the effectiveness of DIP compared with the popular algorithms.



## 1. Introduction

Dimensionality reduction, one of the fundamental problems in face recognition finds a projection that transforms samples from a high dimensional space to a low dimensional subspace for the subsequent classification. It aims to reveal a succinct and effective representation of the distribution of samples in the original space. Over the past few decades, many dimensionality reduction algorithms have been proposed [2,8,11,15,23,24,30,32], since it can be applied to various applications, such as face recognition [25,27], scene analysis [6,26], object categorization [10,33,34] and biometrics [35]. These algorithms can be grouped into two categories: globally linear dimensionality reduction algorithms and manifold learning based dimensionality reduction algorithms. In recently years, there are a kind of piece-wise linear algorithm [12,13]. These methods approximate a nonlinear model by using piece-wise linear sub-models. The algorithms can be treated as a combination of globally linear dimensionality reduction algorithms and manifold learning based dimensionality.

Principal Component Analysis (PCA) [11] and Linear Discriminant Analysis (LDA) [8] are the representative globally linear dimensionality reduction algorithms. PCA is an unsupervised learning algorithm. By maximizing the trace of the total scatter matrix in the projected low dimensional subspace, it can optimally reconstruct Gaussian distributed data. Since it does not utilize the class label information, we do not apply it to classification task in general. LDA is the typical traditional supervised classification algorithm that utilizes the class label information. It is popular in face recognition [3] and other pattern classification tasks [17]. However, traditional LDA has the following drawbacks. First, it ignores the local structure of samples, so it fails to discover the nonlinear structure hidden in the high dimensional data. Second, LDA is confronted with the small sample size (SSS) problem [22]. A good model approximation usually requires a large number of training samples. Third, since the scatter matrix is almost always singular, it suffers from ill-posed problem for computing the projection matrix in LDA.

Manifold learning based dimensionality reduction algorithms aim to find the intrinsic structure of a set of samples embedded in a high dimensional ambient space [2,14,19,23,29,32]. Laplacian eigenmaps (LE) [2] is a popular unsupervised non-linear classification algorithm. It builds an undirected weighted graph to preserve the sample relationship information in terms of distance. The undirected weighted graph incorporates the neighborhood information of pairwise samples in the training set. LE and other important unsupervised learning algorithms suffer from the out of sample problem [4], e.g., ISOMAP [29] and locally linear embedding (LLE) [23]. They are difficult to find low dimensional embeddings of new test samples. Locality preserving projection (LPP) [14] is the linear approximation of LE and can compute the low dimensional embeddings for new test points. LPP, however, has one important problem: it does not utilize the class label information, so it is not optimal for classification tasks.

Marginal Fisher's Analysis (MFA) [30] is a supervised manifold learning algorithm. By using the inter-class marginal samples, MFA conducts the penalty graph to keep the separability of the inter-class. However, MFA ignores the discriminative information of non-marginal samples and faces the ill-posed problem.

Discriminative locality alignment (DLA) [32] is a new supervised manifold learning algorithm. DLA operates in two main stages to

* Corresponding author. Tel.: +86 20 87113540.
*E-mail addresses:* dapeng.tao@gmail.com (D. Tao), lianwen.jin@gmail.com (L. Jin).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.02.024

overcome the above problems of LDA. In the first stage, DLA aims to preserve the discriminative information in a local patch through the classification optimization criteria that the distance between the intra-class samples will be as small as possible and the distance between the inter-class samples will be as large as possible. In the second stage, DLA integrate all the weighted part optimization to form a global subspace structure. However, DLA is not well suit to preservation of the pairwise measurements in the intra-class samples.

Patch Alignment Framework (PAF) [31] was proposed to understand the common thought and essential difference of these different algorithms. Under the framework of PAF, existing algorithms can be divided into two steps, part optimization and whole alignment. The step of part optimization reveals the intrinsic differences of these algorithms and the step of whole alignment is almost identical in all the dimensionality reduction algorithms.

In this paper, we present Discriminative Information Preservation (DIP) that is a new dimensionality reduction algorithm developed under the framework of PAF. It not only overcomes all the aforementioned problems in conventional LDA, but also enhances the performance of discriminative information extracted from the local patches. DIP operates in the following two steps. First, DIP models the local geometry of intra-class samples by using LPP rebuilt on PAF. Second, the discriminative information of inter-class samples is modeled by maximizing the margin. DIP has three advantages: (1) it focuses on the local patch constructed by each sample and neighborhood, and therefore the non-linearity of the distribution of samples is well modeled, (2) by maximizing the margin between the inter-class samples on the patch, discriminative information is better preserved than other algorithms, and (3) it does not need to compute the inverse of a matrix, and therefore it dose not have the ill-posed problem. In addition, DIP is extended to the semi-supervised DIP (SDIP) by incorporating the unlabeled samples distribution information.

The rest of the paper is organized as follows: Section 2 details the proposed DIP algorithm and extends DIP to semi-supervised (SDIP). In Section 3, we evaluate DIP and SDIP. Finally, concluding remarks and suggestions for future work are presented in Section 4.

## 2. Discriminative information preservation

Consider the problem of discriminative dimension reduction. Denote a set of training samples in a high dimensional space $R^D$ by $X=[x_1,x_2,\ldots,x_N]\in R^{D\times N}$, each of which has a label $C_i\in Z^n$. The objective of discriminative dimension reduction is to find a linear mapping $U\in R^{D\times d}$ from the high dimensional space $R^D$ to a low dimensional subspace $R^d$, with $d<D$. Thus, we can obtain the corresponding low dimensional representation $Y=U^TX=[y_1, y_2,\ldots,y_N]\in R^{d\times N}$.

Under the framework of patch alignment framework (PAF), we present a new discriminative dimension reduction tool, namely, Discriminative Information Preservation (DIP). PAF offers a convenient way to encode the local geometric information of the intra-class samples and discriminative information of the inter-class samples. DIP models a new structure that preserves the pairwise relationship of the intra-class samples by using the classification optimization criterion and transfers this structure from local coordinate system to an aligned coordinate system ruled by PAF. The pairwise relationship in DIP refers to an integration of local geometry and discriminative information.

### 2.1. Part optimization

Given a sample $x_i$, we can divide its $K(=k_1+k_2)$ nearest neighbor samples into two groups. The first group $x_{i^1},x_{i^2},\ldots,x_{i^{k_1}}$ contains $k_1$ samples in the same class with respect to $x_i$, i.e., the intra-class samples. The second group $x_{i_1},x_{i_2},\ldots,x_{i_{k_2}}$ contains $k_2$ samples from different classes with respect to $x_i$, i.e., the inter-class samples. Thus, the local patch for $x_i$ is $X_i=[x_i,x_{i^1},x_{i^2},\ldots,x_{i^{k_1}}, x_{i_1},x_{i_2},\ldots,x_{i_{k_2}}]$. Suppose the corresponding low dimensional representation is $Y_i=[y_i,y_{i^1},y_{i^2},\ldots,y_{i^{k_1}},y_{i_1},y_{i_2},\ldots,y_{i_{k_2}}]$, according to PAF, the part optimization can be formulated as

$$\arg\min_{Y_i} \mathrm{tr}(Y_iL_iY_i^T) \tag{1}$$

where $L_i\in R^{(K+1)(K+1)}$ implies a particular optimization criterion.

In order to encode the local geometry of intra-class samples and the discriminative information of inter-class samples, DIP characterizes three specific properties into $L_i$: (1) the intra-class samples in the high-dimensional space are close to each other in the learned low dimensional space, (2) the inter-class samples are well separable in the learned low-dimensional space, and (3) a trade-off parameter balances the two types of information.

#### 2.1.1. Local geometry preservation (LGP)

The intra-class local geometry is effective for classification and we use LPP to preserve such information. For patch $X_i$, we use $Loc(y_i)$ to define the intra-class local geometry, i.e.,

$$Loc(y_i)=\sum_{j=1}^{k1}\|y_i-y_{i^j}\|^2(w_i)_j \tag{2}$$

where $(w_i)_j=\exp(-\|x_i-x_{i^j}\|^2/t)$ if $x_i\in N_p(x_{i^j})$, otherwise 0, according to Rosenberg [21]. We can also choose $(w_i)_j=1$ if $x_i\in N_p(x_{i^j})$.

In order to use PAF to encode the above LPP based local geometry, we rewrite (2) as

$$\begin{aligned}
Loc(y_i)&=\sum_{j=1}^{k1}\|y_i-y_{i^j}\|^2(w_i)_j\\
&=\mathrm{tr}\left\{\begin{bmatrix}(y_i-y_{i^1})^T\\ \vdots\\ (y_i-y_{i^j})^T\end{bmatrix}[y_i-y_{i^1},\ldots,y_i-y_{i^j}]\mathrm{diag}(w_i)\right\}\\
&=\mathrm{tr}\left\{[y_i-y_{i^1},\ldots,y_i-y_{i^j}]\mathrm{diag}(w_i)\begin{bmatrix}(y_i-y_{i^1})^T\\ \vdots\\ (y_i-y_{i^j})^T\end{bmatrix}\right\}\\
&=\mathrm{tr}\left\{Y_{Loc(i)}\begin{bmatrix}-e_{k1}^T\\ I_{k1}\end{bmatrix}\mathrm{diag}(w_i)[-e_{k1}\quad I_{k1}]Y_{Loc(i)}{}^T\right\}\\
&=\mathrm{tr}(Y_{Loc(i)}L_{Loc(i)}Y_{Loc(i)}^T),
\end{aligned} \tag{3}$$

where $\mathrm{tr}(\cdots)$is the trace operator, $e_{k1}=[1,\ldots,1]^T\in R^{k1}$,

$I_{k1}=\mathrm{diag}(\overbrace{1,\ldots,1}^{k1})$,

$Y_{Loc(i)}=[y_i,y_{i^1},\ldots,y_{i^{k1}}]$

and

$$L_{Loc(i)}=\begin{bmatrix}-e_{k1}^T\\ I_{k1}\end{bmatrix}\mathrm{diag}(w_i)[-e_{k1}\quad I_{k1}]\in R^{(k1+1)(k1+1)}.$$

#### 2.1.2. Discriminative information preservation (DIP)

To preserve the discriminative information for each patch $X_i$, we maximize the margin that is the average difference between the center of the intra-class samples and the inter-class samples, which is defined by

$$Mar'(y_i)=\frac{1}{k_2}\sum_{p=1}^{k_2}\left\|\frac{1}{k_1+1}(y_i+\sum_{j=1}^{k_1}y_{i^j})-y_{i_p}\right\|$$

$$= \frac{1}{k_2}\left\| \frac{k_2}{k_1+1}(y_i + \sum_{j=1}^{k_1} y_{i^j}) - \sum_{p=1}^{k_2} y_{i_p} \right\|$$
$$= \left\| \frac{1}{k_1+1}(y_i + \sum_{j=1}^{k_1} y_{i^j}) - \frac{1}{k_2}\sum_{p=1}^{k_2} y_{i_p} \right\|. \quad (4)$$

In order to use PAF to encode the above margin maximization based discriminative information, we redefine (4) as

$$Mar(y_i) = \left\| \frac{1}{k_1+1}\left(y_i + \sum_{j=1}^{k_1} y_{i^j}\right) - \frac{1}{k_2}\sum_{p=1}^{k_2} y_{i_p} \right\|^2$$
$$= \mathrm{tr}(Y_{Mar(i)} v_i v_i^T Y_{Mar(i)}^T) = \mathrm{tr}(Y_{Mar(i)} L_{Mar(i)} Y_{Mar(i)}^T), \quad (5)$$

wherein

$$v_i = \left[ \overbrace{1/(k_1+1), \ldots, 1/(k_1+1)}^{k_1+1}, \overbrace{-1/k_2, \ldots, -1/k_2}^{k_2} \right]^T,$$

$$Y_{Mar(i)} = [y_i, y_{i^1}, \ldots, y_{i^{k1}}, y_{i_1}, \ldots, y_{i_{k2}}],$$

and $L_{Mar(i)}(i) = v_i v_i^T \in R^{(k1+k2+1)(k1+k2+1)}$.

### 2.1.3. Information fusion

By combining (2) and (4) via a trade-off parameter $\gamma$, we can encode both the local geometry and the discriminative information, i.e.,

$$\arg\min_{y_i}(Loc(y_i) - \gamma Mar(y_i)). \quad (6)$$

After plugging (3) and (5), the objective function (6) turns to

$$\arg\min_{Y_i}(\mathrm{tr}(Y_{Loc(i)} L_{Loc(i)} Y_{Loc(i)}^T) - \gamma\, \mathrm{tr}(Y_{Mar(i)} L_{Mar(i)} Y_{Mar(i)}^T)). \quad (7)$$

## 2.2. Whole alignment

Suppose the low-dimensional representation of the patch $X_i$ is $Y_i$. According to PAF, we will unify all the low dimensional representations $Y_i$ into a consistent coordinate. The coordinate of $Y_i$ is selected from the global coordinate $Y=[y_1, y_2, \ldots, y_n] \in R^{d \times N}$. This process requires a selection matrix $S_i$ that selects samples to form the patch $Y_i=[y_i, y_{i1}, \ldots, y_{iK}] \in R^{d(K+1)}$, i.e.,

$$Y_i = YS_i. \quad (8)$$

Let $F_i = \{i, i_1, \ldots, i_K\}$ be the index set. The selection matrix $S_i$ is defined by

$$(S_i)_{pq} = \begin{cases} 1, & \text{If } p = F_i\{q\} \\ 0, & \text{else} \end{cases} \quad (9)$$

We define two different selection matrices $S_{Loc(i)}$ whose index set is $F_i = \{i, i^1, \ldots, i^{k_1}\}$, and $S_{Mar(i)}$ whose index set is $F_i = \{i, i^1, i^2, \ldots, i^{k_1}, i_1, i_2, \ldots, i_{k_2}\}$. Thus, $Y_{Loc(i)}$ and $Y_{Mar(i)}$ can be rewritten as

$$Y_{Loc(i)} = YS_{Loc(i)} \quad (10)$$

and

$$Y_{Mar(i)} = YS_{Mar(i)}, \quad (11)$$

respectively.

According to (10) and (11), the part optimization (7) can be rewritten as

$$\arg\min_{Y}(\mathrm{tr}(YS_{Loc(i)} L_{Loc(i)} S_{Loc(i)}^T Y^T) - \gamma \mathrm{tr}(YS_{Mar(i)} L_{Mar(i)} S_{Mar(i)}^T Y^T)). \quad (12)$$

By summing over all the part optimizations defined in (12), we have the whole alignment

$$\arg\min_{Y} \sum_{i=1}^{N} (\mathrm{tr}(YS_{Loc(i)} L_{Loc(i)} S_{Loc(i)}^T Y^T) - \gamma \mathrm{tr}(YS_{Mar(i)} L_{Mar(i)} S_{Mar(i)}^T Y^T)). \quad (13)$$

It can be simplified according to

$$\arg\min_{Y} \sum_{i=1}^{N} (\mathrm{tr}(YS_{Loc(i)} L_{Loc(i)} S_{Loc(i)}^T Y^T) - \gamma \mathrm{tr}(YS_{Mar(i)} L_{Mar(i)} S_{Mar(i)}^T Y^T))$$
$$= \arg\min_{Y} \mathrm{tr}\left( Y \left( \sum_{i=1}^{N} (S_{Loc(i)} L_{Loc(i)} S_{Loc(i)}^T) \right) Y^T \right)$$
$$- \mathrm{tr}\left( Y \left( \gamma \sum_{i=1}^{N} (S_{Mar(i)} L_{Mar(i)} S_{Mar(i)}^T) \right) Y^T \right)$$
$$= \arg\min_{Y} \mathrm{tr}\left( Y \left( \sum_{i=1}^{N} (S_{Loc(i)} L_{Loc(i)} S_{Loc(i)}^T) \right.\right.$$
$$\left.\left. - \gamma \sum_{i=1}^{N} (S_{Mar(i)} L_{Mar(i)} S_{Mar(i)}^T) \right) Y^T \right)$$
$$= \arg\min_{Y} \mathrm{tr}(YLY^T), \quad (14)$$

where

$$L = \sum_{i=1}^{N} (S_{Loc(i)} L_{Loc(i)} S_{Loc(i)}^T) - \gamma \sum_{i=1}^{N} (S_{Mar(i)} L_{Mar(i)} S_{Mar(i)}^T) \in R^{N \times N}$$

is the whole alignment matrix encoding both the local geometry of the intra-class samples and the discriminative information of the inter-class samples. If the initialization of $L$ is set to zero, we can obtain it by using an iterative procedure

$$L(F_i, F_i) \leftarrow L(F_i, F_i) + L_{Loc(i)} - \gamma L_{Mar(i)}, \quad (15)$$

where $i = 1, \ldots, N$. It is worth to note that the calculations of (3) and (5) show the whole alignment matrix is symmetric.

## 2.3. Low dimensional embedding

To uniquely determine $Y$, we impose $YY^T = I_d$ on (14), wherein $I_d$ is a $d \times d$ identity matrix, i.e.,

$$\arg\min_{Y} (YLY^T)$$
$$\text{s.t. } YY^T = I_d. \quad (16)$$

By using the Lagrange multiplier method [16], (16) can be transformed to a generalized eigenvalue problem and thus $Y$ is formed by $d$ eigenvectors associated with $d$ smallest eigenvalues of $L$.

## 2.4. Linear approximation

The above method suffers from the out of sample problem [4]. By applying linearization, we can explore an explicit embedding for a sample. For linearization, we can simply impose a constraint $U^TU = I_d$ on (14) to determine the projection matrix $U$ according to $Y = U^TX$, wherein $I_d$ is a $d \times d$ identity matrix. Thus (14) is transformed to (17)

$$\arg\min_{Y} \mathrm{tr}(U^T XLX^T U)$$
$$\text{s.t. } U^T U = I_d. \quad (17)$$

Similar to (16), it can be transformed to a generalized eigenvalue problem and $U$ is given by $d$ eigenvectors associated with $d$ smallest eigenvalues of $XLX^T$. In practice, PCA can be applied to the original high dimensional data for removing the noise. The main steps of DIP with linearization are summarized in Table 1.

## 2.5. Semi-supervised Discriminative Information Preserving (SDIP)

It has been widely acknowledged that unlabeled samples are useful to enhance the classification performance [5]. In practice, it is possible to collect a large number of unlabeled samples and

**Table 1**
DIP with linearization.

| Algorithm: Linear Discriminative Information Preserving (LDIP) | |
|---|---|
| Input: | Training set $X=[x_1,x_2,...,x_n]\in R^{D\times N}$;<br>Class label vector $C=[c_1,c_2,\ldots,c_n]^T$;<br>$d$: dimension of the reduced space. |
| Output: | Linear projection matrix $U=[u_1,u_2,\ldots,u_d]\in R^{D\times d}$ |
| Step1: | Optional PCA reconstruction of original training set $X$, and the PCA projection matrix is $U_{PCA}$; |
| Step 2: | Part optimization: construct $N$ patches for the training set according to the models of LGP and DIP, calculate the matrixes $L_{Loc(i)}$ and $L_{Mar(i)}$ for each patch using (3) and (5); |
| Step 3: | Whole alignment: sum over all the patches in a global coordinate, computing the whole alignment matrix $L$ using (15); |
| Step 4: | Compute project matrix $U_{DIP}$ whose column vectors are the $d$ eigenvectors of $XLX^T$ associated with $d$ smallest eigenvalues. |
| Step 5: | Return the final projection matrix $U=UPCAUDIP$. |

the sample distribution can be deemed as a prior to improve the decision making. Therefore, by taking unlabeled samples into account, we improve DIP by proposing a semi-supervised extension, or semi-supervised DIP (SDIP) which preserves the local geometry of both the labeled and unlabeled samples. Similar to DIP, SDIP maximizes the margin of different classes.

We attach the unlabeled samples to the original data set, and thus we have $X=[x_1,\ldots,x_N,x_{N+1},\ldots,x_{N+N_U}]\in R^{D(N+N_U)}$, wherein the first $N$ samples are labeled, and the rest $N_U$ samples are unlabeled. The modeling of the labeled samples in SDIP is the same as that in DIP. In SDIP, the neighborhood connection between each unlabeled sample $x_i, i=N+1,\ldots,N+N_U$ and its nearest neighbors $x_{i_1},...,x_{i_{ks}}$ are expected to be remained in the low dimensional subspace. Let $X_i=[x_i,x_{i_1},\ldots,x_{i_{ks}}]$ denote the $i$th patch of the unlabeled samples, and the affiliation index set is $F_i^U=\{i,i_1,\ldots,i_{ks}\}$. Therefore, the part optimization for the unlabeled samples is defined by

$$\arg\min_{y_i}\sum_{j=1}^{ks}\|y_i-y_{i_j}\|^2(o_i)_j, \tag{18}$$

where $y_{i_j}, j=1,\ldots,ks$, are $ks$ neighbor samples including both the labeled and the unlabeled samples in the local patch; and $(o_i)_j$ is the weighting vector calculated by the heat kernel $(o_i)_j=\exp(-\|x_i-x_{i_j}\|^2/t)$.

Similar to (3), (18) can be rewritten as

$$\begin{aligned}Un(y_i)&=\sum_{j=1}^{ks}\|y_i-y_{i^j}\|^2(o_i)_j\\&=\mathrm{tr}\left\{Y_{Un(i)}\begin{bmatrix}-e_{ks}^T\\ I_{ks}\end{bmatrix}\mathrm{diag}(o_i)[-e_{ks}\quad I_{ks}]{Y_{Un(i)}}^T\right\}\\&=\mathrm{tr}(Y_{Un(i)}L_{Un(i)}Y_{Un(i)}^T),\end{aligned} \tag{19}$$

where

$$L_{Un(i)}=\begin{bmatrix}-e_{ks}^T\\ I_{ks}\end{bmatrix}\mathrm{diag}(o_i)[-e_{ks}\quad I_{ks}]\in R^{(ks+1)(ks+1)},$$

$e_{ks}=[1,\ldots,1]^T\in R^{ks}$, and $I_{k1}=\mathrm{diag}\left(\overbrace{1,\ldots,1}^{ks}\right)$

It can be treated as a regularization item for modeling the data distribution prior.

Since the number of labeled samples is usually much smaller than that of unlabeled samples, we only use unlabeled samples to form the regularizer for learning the marginal distribution.

Therefore, SDIP can be written as

$$\begin{aligned}\arg\min_{Y_i}&\left(\sum_{i=1}^{N}\begin{pmatrix}\mathrm{tr}(Y_{Loc(i)}L_{Loc(i)}Y_{Loc(i)}^T)\\ -\gamma\,\mathrm{tr}(Y_{Mar(i)}L_{Mar(i)}Y_{Mar(i)}^T)\end{pmatrix}\right.\\&\left.+\beta\sum_{i=N+1}^{N+N_U}\mathrm{tr}(Y_{Un(i)}L_{Un(i)}Y_{Un(i)}^T)\right).\end{aligned} \tag{20}$$

We define a selection matrix $S_{Un(i)}$ for unlabeled samples and rewritten $Y_{Un(i)}$ as

$$Y_{Un(i)}=YS_{Un(i)}. \tag{21}$$

Therefore, (20) can be rewritten as

$$\begin{aligned}&\arg\min_{Y}\left(\sum_{i=1}^{N}\begin{pmatrix}\mathrm{tr}(YS_{Loc(i)}L_{Loc(i)}S_{Loc(i)}^TY^T)\\ -\gamma\mathrm{tr}(YS_{Mar(i)}L_{Mar(i)}S_{Mar(i)}^TY^T)\end{pmatrix}\right.\\&\quad\left.+\beta\sum_{i=N+1}^{N+N_U}\mathrm{tr}(YS_{Un(i)}L_{Un(i)}S_{Un(i)}^TY^T)\right)\\&=\arg\min_{Y}\mathrm{tr}\left(Y\left(\sum_{i=1}^{N}\begin{pmatrix}S_{Loc(i)}L_{Loc(i)}S_{Loc(i)}^T\\ -\gamma S_{Mar(i)}L_{Mar(i)}S_{Mar(i)}^T\end{pmatrix}\right.\right.\\&\quad\left.\left.+\sum_{i=N+1}^{N+N_U}\beta S_{Un(i)}L_{Un(i)}S_{Un(i)}^T\right)Y^T\right)\\&=\arg\min_{Y}\mathrm{tr}(YL^SY),\end{aligned} \tag{22}$$

where

$$\begin{aligned}L^S=&\left(\sum_{i=1}^{N}\begin{pmatrix}S_{Loc(i)}L_{Loc(i)}S_{Loc(i)}^T\\ -\gamma S_{Mar(i)}L_{Mar(i)}S_{Mar(i)}^T\end{pmatrix}\right.\\&\left.+\sum_{i=N+1}^{N+N_U}\beta S_{Un(i)}L_{Un(i)}S_{Un(i)}^T\right)\in R^{(N+N_U)(N+N_U)}\end{aligned}$$

is the whole alignment matrix of SDIP, and $\beta$ is a trade-off parameter to balance the supervised information and the unsupervised information.

This $L^S$ can be updated according to

$$\begin{cases}L^S(F_i,F_i)\leftarrow L^S(F_i,F_i)+L_{Loc(i)}-\gamma L_{Mar(i)} & \text{for } i=1,\ldots,N,\\ L^S(F_i^U,F_i^U i)\leftarrow L^S(F_i^U,F_i^U)+\beta L_{Un(i)} & \text{for } i=N+1,\ldots,N+N_U,\end{cases} \tag{23}$$

and the initialization of $L^S$ is set to zero.

SDIP can also be solved by generalized singular value decomposition and the linearization can be obtained in accordance with (17).

### 2.6. Time complexity analysis

Suppose that we are given $N$ training samples in a $D$ dimensional space. The time complexity of DIP contains two parts. One part is for the computation of the whole alignment matrix $L$. The time complexity of this part is $O((D+K)\times N^2)$, where $K$ is the number of nearest neighbor samples. When $K<<D$, we have $O(D\times N^2)$. The other part is for the computation of the eigenvalue problem. The time complexity of this part is $O(N^3)$. Therefore, the whole time complexity of DIP is $O(D\times N^2+N^3)$. The time complexity of SDIP is the same as that of DIP but $N$ refers to the total number of training samples including unlabeled samples.

## 3. Experiments

We evaluate the performance of the proposed DIP with five representative algorithms, including PCA [1,11], LDA [8], SLPP (LPP1 in [7]), MFA [30] and DLA [32]. These algorithms have certain merits in their own rights. PCA and LPP are unsupervised algorithms that do not consider the class label information. LDA, MFA and DLA are all supervised algorithms.

We randomly divided datasets that constructed from each dataset into three separate sets, i.e., validation set, training set and test set. Validation set can be used to tune the optimal parameters in algorithms. For the proposed DIP algorithm, validation set was used to determine the important parameters that include $k_1$ (the number of intra-class samples), $k_2$ (the number of inter-class samples), $\gamma$ (the ratio of information fusion) and $d$ (the subspace dimension). Training set was used to learn the

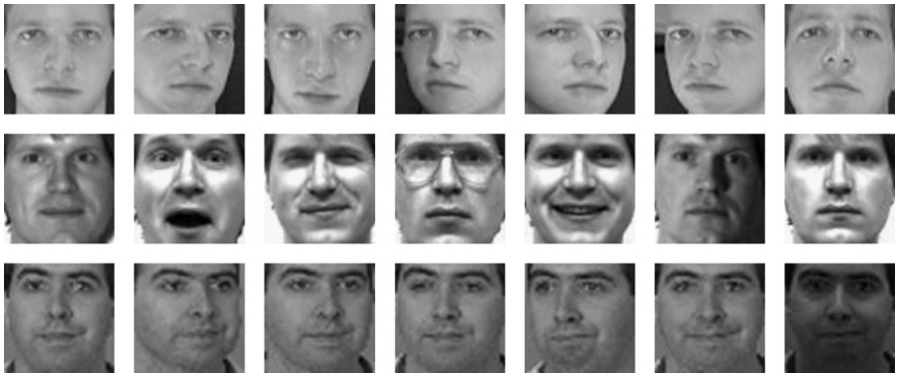

**Fig. 1.** Example face images come from different dataset. The first row selects from ORL; the second row selects from YALE; and the third row selects from FERET.

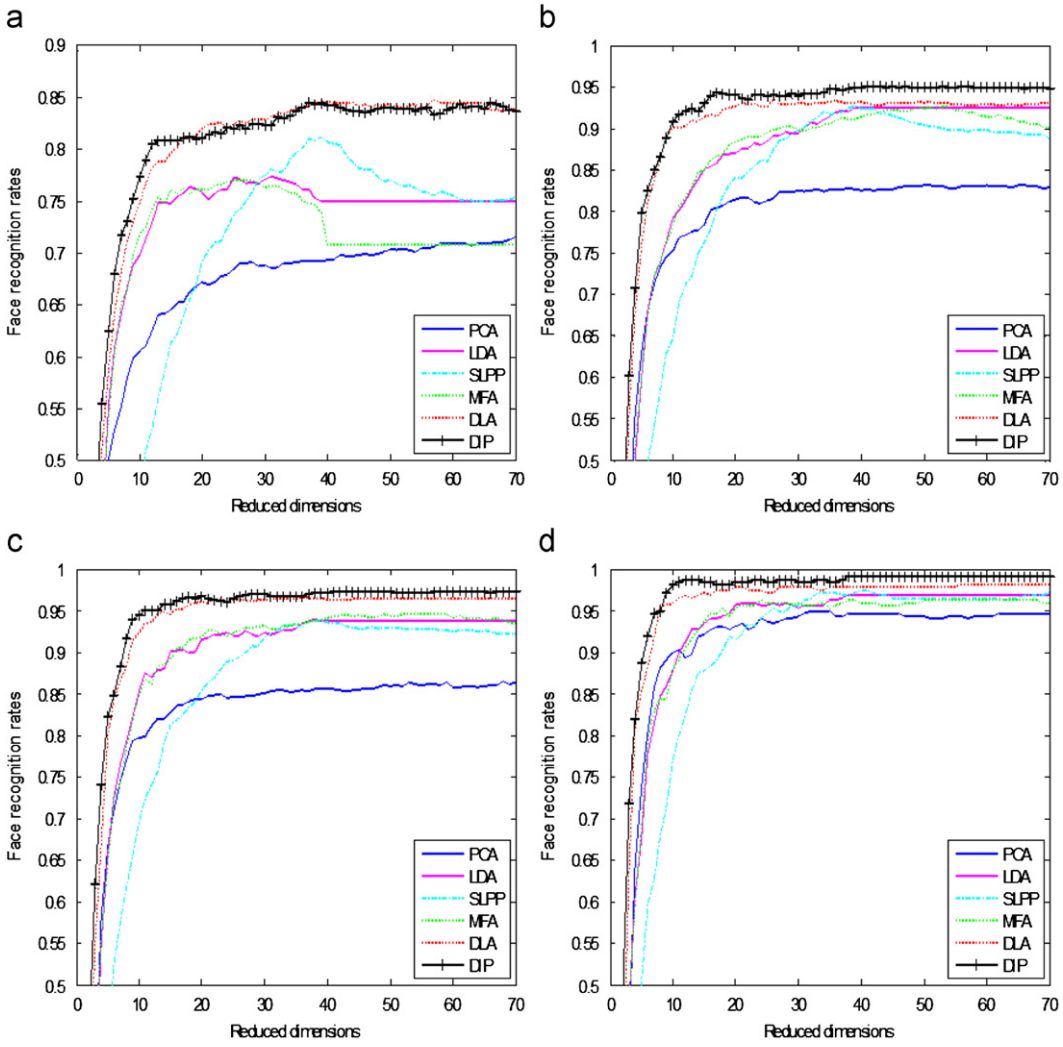


**Fig. 2.** Face recognition rate vs. dimensionality reduction on the validation sets of ORL: (a) two samples for training, (b) four samples for training, (c) six samples for training and (d) eight samples for training.

projection matrix and low-dimensional representation of training samples. The test set was used for performance evaluation. The accuracy is the percentage classified correctly of samples in the test set. We use the Nearest Neighbor (NN) rule in classification during validation and test stages.

Before we conduct LDA, SLPP, MFA, DLA and DIP, we first use PCA to remove redundant information. In the PCA step, $N$–$C$ dimension of samples are retained to ensure that $X(D^p-W^p)X^T$ [30] in MFA and within-scatter matrix $S_w$ in LDA [18] are non-singular for these algorithms. Although LPP, DLA and DIP algorithms have no singularity problem, we first apply PCA projection and subspaces are set $N$–1 dimensions.

We selected three face image datasets that include ORL [28], YALE [9] and FERET [20]. Fig. 1 shows example images of ORL, YALE and FERET datasets. In order to conduct face recognition, we align all face images according to the eye position and linearly rescale each pixel of image to gray level of 256. After all images were normalized to $32 \times 32$ pixel array, we scanned them into a long vector.

We choose the training samples as reference of each class to evaluate performance, since $k$-nearest-neighbor was utilized to classify the test samples. These experiments were independently conducted ten times and we reported the averaged accuracy.

**Table 2**
Best recognition rates (%) of six algorithms on the ORL testing dataset.

| Number of training samples | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| PCA | 69.97(69) | 84.17(63) | 88.75(70) | 94.75(66) |
| LDA | 76.44(30) | 92.21(39) | 95.2(38) | 97.62(39) |
| SLPP | 78.75(39) | 92.0(39) | 94.65(39) | 97.52(44) |
| MFA | 76.34(29) | 92.42(51) | 95.75(61) | 97.0(68) |
| DLA | **83.59(62)** | 94.71(70) | 96.41(30) | 98.13(25) |
| DIP | 83.5(54) | **95.83(38)** | **97.6(62)** | **99.25(63)** |

The number in the parentheses is the reduced dimensions.
The bold values indicate the highest rates obtained corresponding to each column.

### 3.1. ORL dataset

In the Cambridge ORL dataset, there are 400 images collected from 40 individuals. Ten images were sampled from each individual with varying lighting, facial expressions and facial details (glasses/no-glasses). These images were taken in the same dark background color. In the stage of training, we randomly selected different numbers (2, 4, 6, 8) of images from each individual to construct the training set. The rest of the images were divided equally among test set and validation set. Fig. 2 shows the

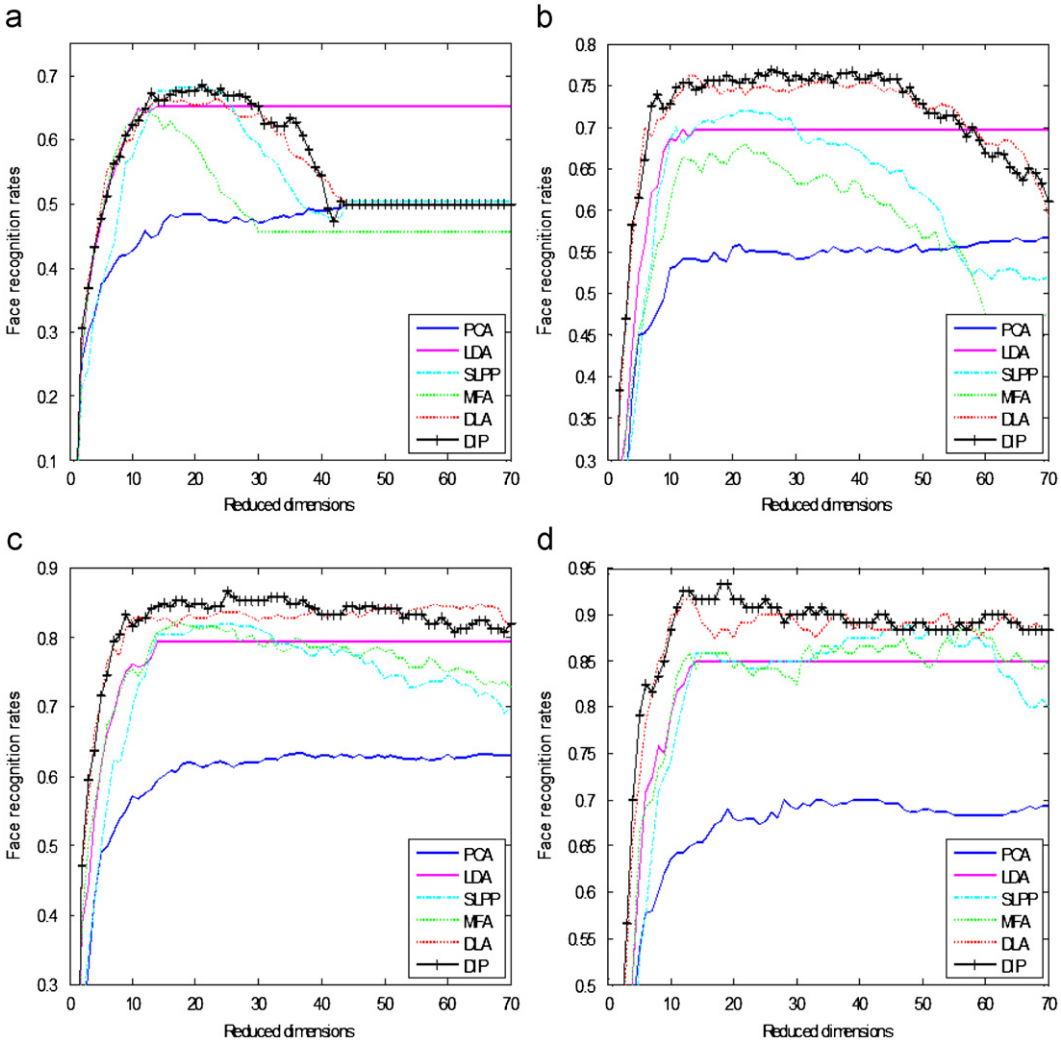


**Fig. 3.** Face recognition rate vs. dimensionality reduction on the validation sets of YALE: (a) three samples for training, (b) five samples for training, (c) seven samples for training and (d) nine samples for training.

average accuracy versus the subspace dimensions on the validation set. Table 2 reports the best accuracy and corresponding dimension of all the algorithms on the test set. Furthermore, Fig. 8(a) shows box-and-whisker plots of DIP and DLA to describe the recognition performances with statistical significance. We observe that DIP algorithm outperforms the others in general except for the situation that the training samples from each individual is 2. Parameters $k_1$ , $k_2$ are important for patch building and the trade-off parameter $\gamma$ is important for information infusion. Section 3.4 shows how to choose these three parameters of DIP.

### 3.2. YALE dataset

The YALE dataset consists of 165 frontal view face images collected from 15 individuals. There are eleven images for each individual with varying facial expressions, or configurations. In the stage of training, we randomly selected different numbers (3, 5, 7, 9) of images from each individual to construct the training set. The rest of the images were divided equally among test set and validation set. Fig. 3 shows the average accuracy versus the subspace dimensions on the validation set. Table 3 reports the best accuracy and the corresponding dimension of all the algorithms on the test set. Fig. 8(b) shows box-and-whisker plots of DIP and DLA to describe the recognition performances with statistical significance. We observe that DIP algorithm outperforms the others.

### 3.3. FERET dataset

The complete FERET dataset contains 13,539 face images collected from 1565 individuals. All the images vary in pose, illumination, size, facial expression and age. In our experiment, 100 individuals, having seven images, were randomly selected. In the stage of training, we randomly selected different numbers (3, 5) of images from each individual to construct the training set. The rest of the images were divided equally among test set and validation set. Fig. 4 shows the average accuracy versus the subspace dimensions on the validation set. Table 4 reports the best accuracy and corresponding dimension of all the algorithms on the test set. In order to report the recognition performances with statistical significance, we perform box-and-whisker plots of DIP and DLA in Fig. 8(c). It can be seen that DIP performs better than other algorithms.

### 3.4. Parameter selection

DIP has three parameters, i.e. patch building parameters $k_1$, $k_2$ and trade-off parameter $\gamma$. Effects of these parameters on the recognition accuracy were studied in this section. In the validation phase, we selected eight samples in each class of the ORL dataset and the Yale dataset. Since there are seven samples in each class of the FERET dataset, we selected five samples in each class of the FERET as validating set. In the experiment, we fixed the selected subspace dimension to 30.

At first, we analyze the effect of the trade-off parameter $\gamma$, by fixing patch building parameters $k_1=3$ and $k_2=1$. In Fig. 5, the correlation between trade-off parameter $\gamma$ and the recognition

**Table 3**
Best recognition rates (%) of six algorithms on the YALE testing dataset.

| Number of training samples | 3 | 5 | 7 | 9 |
|---|---|---|---|---|
| PCA | 49.08(44) | 58.44(70) | 61.0(45) | 67.0(37) |
| LDA | 58.75(14) | 76.11(14) | 79.5(14) | 82.0(14) |
| SLPP | 65.08(13) | 77.78(15) | 81.33(14) | 81.67(14) |
| MFA | 59.42(13) | 72.11(21) | 81.83(23) | 84.33(58) |
| DLA | 63.92(17) | 79.22(31) | 82.67(29) | 86.33(14) |
| DIP | **65.67(18)** | **80.67(38)** | **83.5(14)** | **88.67(14)** |

The number in the parentheses is the reduced dimensions.
The bold values indicate the highest rates obtained corresponding to each column.

**Table 4**
Best recognition rates (%) of six algorithms on the FETET testing dataset.

| Number of training samples | 3 | 5 |
|---|---|---|
| PCA | 40.33(70) | 50.9(68) |
| LDA | 50.2(31) | 56.05(62) |
| SLPP | 41.9(70) | 44.85(69) |
| MFA | 54.67(36) | 52.65(70) |
| DLA | 84.85(31) | 91.45(26) |
| DIP | **85.75(14)** | **92.75(25)** |

The number in the parentheses is the reduced dimensions.
The bold values indicate the highest rates obtained corresponding to each column.

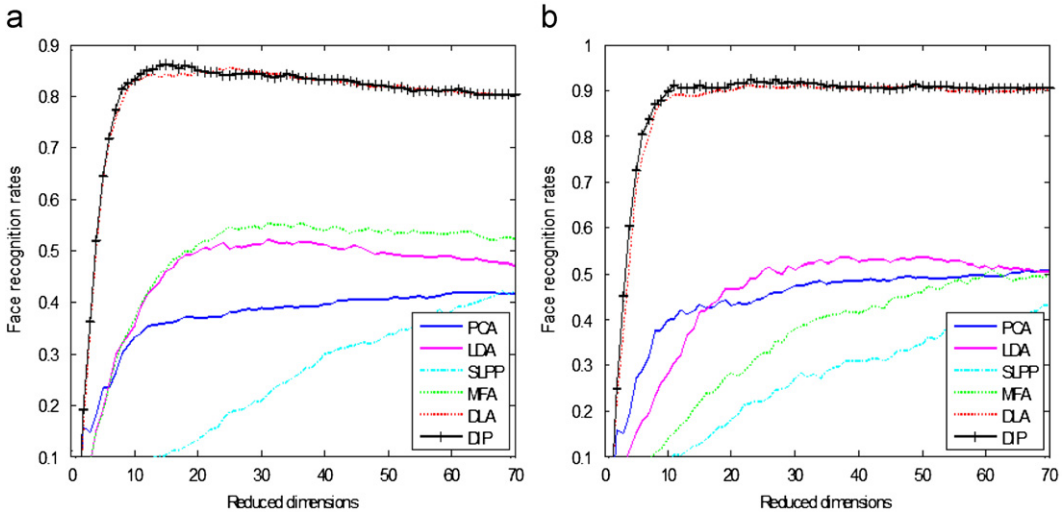


**Fig. 4.** Face recognition rate vs. dimensionality reduction on the validation sets of FERET: (a) three samples for training and (b) five samples for training.

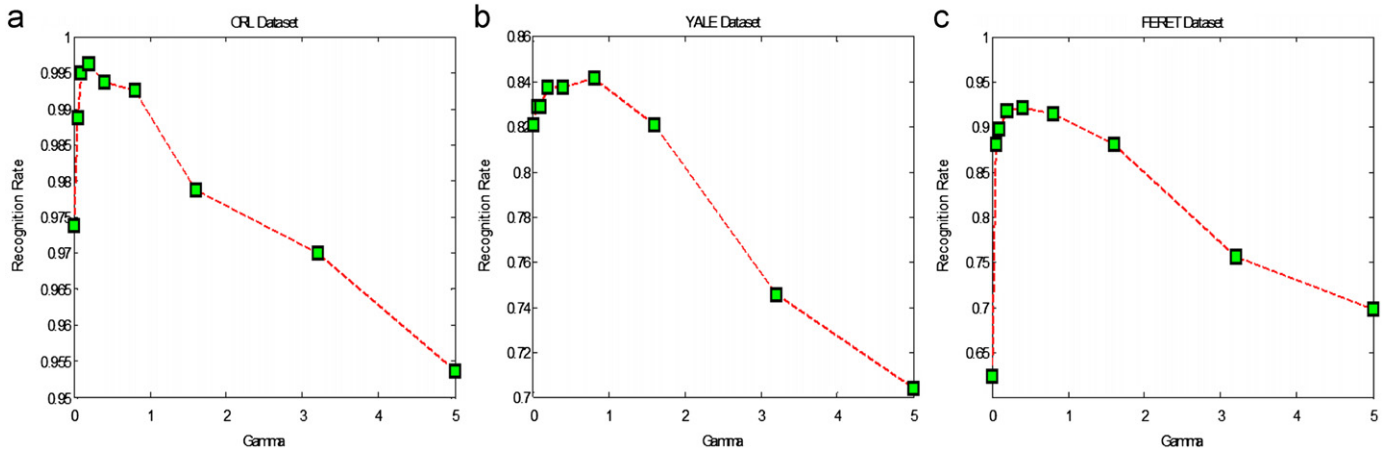


**Fig. 5.** Trade-off parameter vs. face recognition rate.

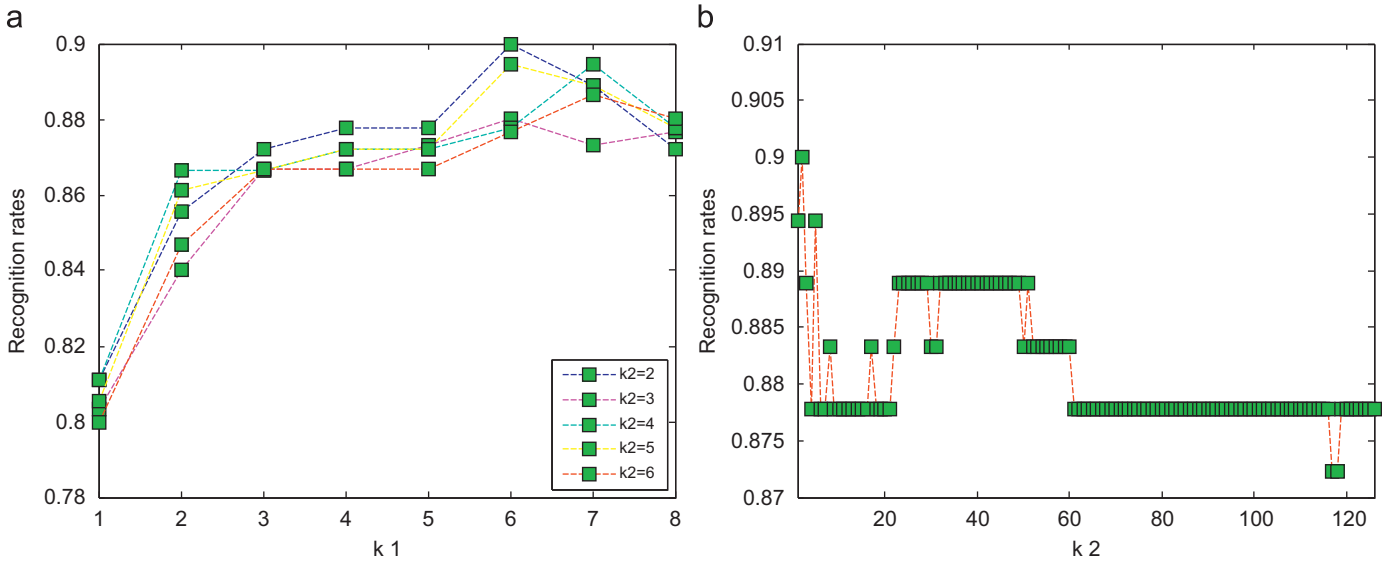


**Fig. 6.** (a) Face recognition rate vs. $k_1$ and (b) face recognition rate vs. $k_2$.

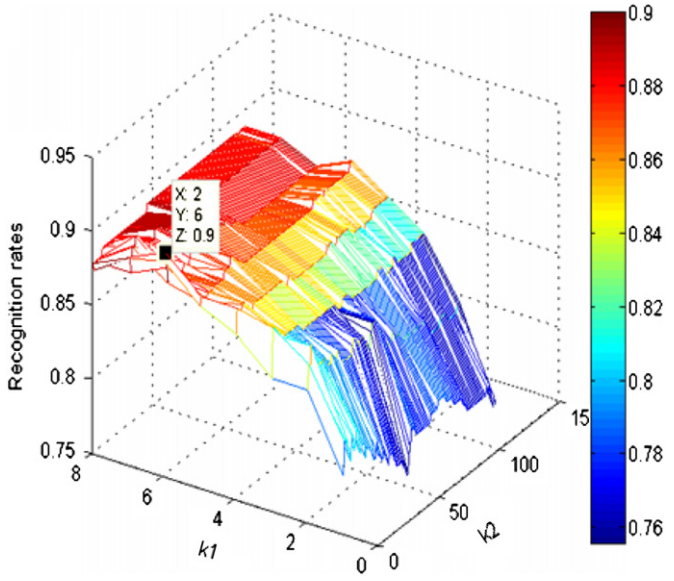


**Fig. 7.** Face recognition rate vs. $k_1$ and $k_2$.

rate of the face recognition are shown. Base on the figures, we observe that the best face recognition rates are obtained when $\gamma$ is neither too small nor too large. This reveals the optimal face recognition rate is a balance between the local geometry and the discriminative information. Therefore, we can conclude that by using a proper trade-off parameter, the local geometry and the discriminative informative can be effectively fused.

Secondly, we analyze the effects of patch building parameters $k_1$ (the number of neighbor samples of intra-class) and $k_2$ (the number of neighbor samples of inter-class) on the face recognition rates based on the YALE dataset, by fixing trade-off parameter $\gamma=1$. Nine samples in each class were selected in the validation stage, the rest were used for test. We fixed the selected subspace dimension to 30.

Suppose $N_i$ is the training sample number in each class, we vary $k_1$ from 1 to $N_i$–1(=8) by fixing $k_2$ to an arbitrary value that is not greater than $N$–$N_i$ in order to achieve the face recognition rate curve. Fig. 6(a) shows face recognition rate curve with respect to $k_1$. By fixing $k_2=2$, the peak of curve can be acquired when $k_1=6$. We vary $k_2$ from 1 to $N$–$N_i$(=126) by fixing $k_1=6$ to achieve another face recognition rate curve. Fig. 6(b) shows face recognition rate curve with respect to $k_2$. The peak of curve can be acquired when $k_2=2$. We also vary $k_1$ from 1 to 8 and $k_2$ from 1 to 126 simultaneously. Fig. 7 shows that best face recognition rate with the corresponding

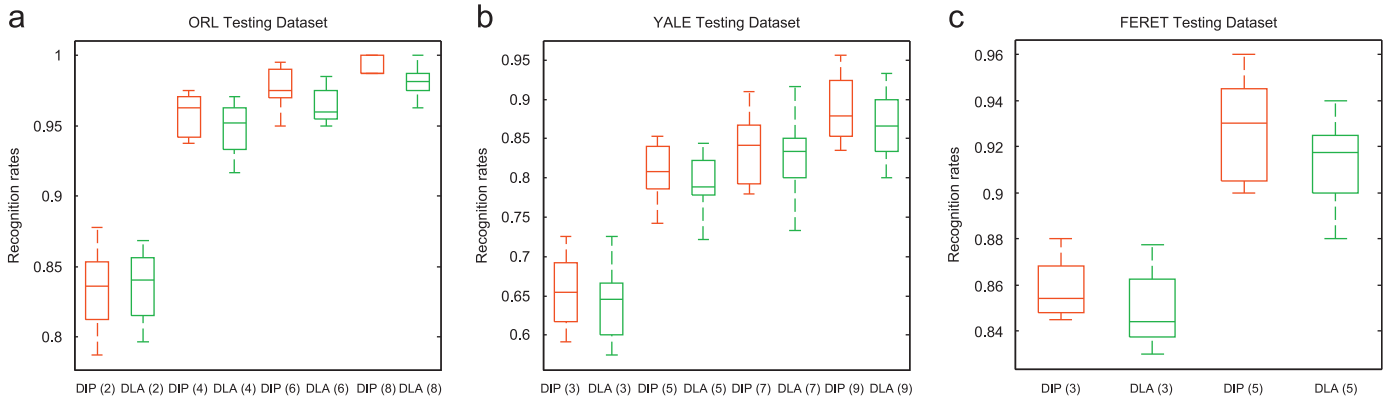


**Fig. 8.** Recognition performances with statistical significance. Note that the number in the parentheses is number of training samples from each class.

**Table 5**
Best recognition rates (%) of DIP and SDIP on the YALE dataset.

| | 2 labeled | 4 labeled |
|---|---|---|
| 3 unlabeled | | |
| DIP | 55.37(12) | 73.97(21) |
| SDIP | **56.22(18)** | **75.12(30)** |
| 5 unlabeled | | |
| DIP | 55.56(16) | 74.27(20) |
| SDIP | **57.16(17)** | **76.0(14)** |

The number in the parentheses is the reduced dimensions.
The bold values indicate the highest rates obtained corresponding to each column.

$k_1=6$ and $k_2=2$ in this experiment. Since patch building parameters $k_1$ and $k_2$ suggest that the algorithm models the range of local neighborhood, therefore, we can conclude that DIP can achieve better performance in the local neighborhood other than the global structure.

### 3.5. *Semi-supervised experiments*

We compare SDIP and DIP based on the YALE dataset. We independently conducted the experiment ten times and reported the averaged accuracy. In the stage of training, we randomly selected different numbers (2, 4) of images with labels and different numbers (3, 5) of images without labels from each individual to construct the training set. The rest of image build of test set for the test stage. In fact, the number of training samples without labels has no effects in training DIP. It can be concluded from Table 5 that unlabeled samples are useful to improve recognition rates.

### 3.6. *Discussion*

We have a number of interesting points based on the experiments above:

1. DIP models both the local geometric information and discriminative information. Therefore, it works better than other algorithms such as PCA, LDA, SLPP, MFA and DLA. We note that although MFA considers both two aspects, its recognition rate is not as good as DIP. Because PCA used in MFA for data preprocessing discards useful discriminative information and MFA ignores the discriminative information carried by non-marginal samples of each class. The recognition rate achieved by DLA is lower than that achieved by DIP, because DLA does not properly handle the pairwise measurements in the intra-class samples.
2. Fig. 3(a) and (b) show that the decline in recognition rate emerges with the increase of the dimension of DIP subspace, when a classification optimal dimension of DIP subspace is achieved. This phenomenon can be explained by the reason that the training set size is much smaller than the dimension of sample.
3. Figs. 5–7 show that the performance of DIP is robust to wide ranges of the trade-off parameter $\gamma$, the number of intra-class samples $k_1$ and the number of inter-class samples $k_2$ in our experiments.
4. It is possible to extend DLA and LPP to their semi-supervised versions for modeling the marginal distribution of the samples. However, it is direct to foresee that semi-DIP outperforms semi-DLA and semi-LPP because DIP outperforms DLA and LPP, because DIP enjoys several advantages compared with DLA and LPP.

## 4. Conclusion

This paper presented a new dimensionality reduction algorithm, termed Discriminative Information Preservation (DIP), under the framework of PAF. In contrast to other popular manifold learning algorithms, DIP preserves both the local geometry of the intra-class samples and the discriminative information of the inter-class samples in part optimization; and avoids computing the inverse of a matrix. Experiments on face recognition show the effectiveness of DIP by comparing with popular dimensionality reduction algorithms, e.g., PCA, LDA, SLPP, MFA, DLA. In the future, we will apply the proposed DIP algorithm to other applications, e.g., image annotation, object classification, scene analysis. Since DIP has three important parameters that determine the algorithm performance, therefore, automatic selection of the parameters is urgent to deeply study.

## Acknowledgment

This work is supported in part by NSFC (Grant nos. U0735004 and 61075021), GDSTP (nos. 2010B090400397 and S2011020000541).

## References

[1] P. Belhumeour, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inform. Process. Syst. (2002) 585–591.
[3] X. Gao, J. Zhong, D. Tao, X. Li, Local face sketch synthesis learning, Neurocomputing 71 (10-12) (2008) 1921–1930.
[4] Y. Bengio, J. Paiement, P. Vincent, O. Dellallaeu, L. Roux, M. Quimet, Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering, Adv. Neural Inform. Process. Syst. (2004).
[5] D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–7.

[6] X. Gao, Z. Niu, D. Tao, X. Li, Non-goal scene analysis for soccer video, Neurocomputing 74 (4) (2011) 540–548.
[7] D. Cai, X. He, J, Han, Using Graph Model for Face Analysis, Department of Computer Science Technical Report no. 2636, University of Illinois at Urbana-Champaign, September 2005.
[8] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188.
[9] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many, illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[10] T. Xia, D. Tao, T. Mei, Y. Zhang, Multiview spectral embedding, IEEE Trans. Syst. Man Cybern. Part B 40 (6) (2010) 1438–1446.
[11] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.
[12] X. Wang, Z. Li, D. Tao, Subspaces indexing model on Grassmann manifold for image search, IEEE Trans. Image Process. 20 (9) (2011) 2627–2635.
[13] Z. Li, Y. Fu, J. Yuan, T.S. Huang, Y. Wu, Query driven local linear discriminant models for head pose estimation, in: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME), 2007.
[14] X. He, P. Niyogi, Locality preserving projections, Adv. Neural Inform. Process. Syst. (2004).
[15] N. Guan, D. Tao, Z. Luo, B. Yuan, Non-negative patch alignment framework, IEEE Trans. Neural Networks 22 (8) (2011) 1218–1230.
[16] I.T. Jolliffe, Principal Component Analysis, 2nd ed, Springer-Verlag, New York, 2002.
[17] L. Jin, K. Ding, Z. Huang, Incremental learning of LDA model for Chinese writer adaptation, Neural Comput. 73 (2010) 1614–1623.
[18] J. Lu, K. Plataniotis, A. Venetsanopoulos, Face recognition using LDA-based algorithms, IEEE Trans. Neural Networks 14 (1) (2003) 195–200.
[19] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, IEEE Trans. Image Process. 20 (7) (2011) 2030–2048.
[20] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[21] S. Rosenberg, The Laplacian on a Riemannian Manifold, Cambridge University Press, 1997.
[22] D. Tao, X. Li, X. Wu, S.J. Maybank, Geometric mean for subspace selection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 260–274.
[23] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[24] X. Gao, X. Wang, D. Tao, X. Li, Supervised Gaussian process latent variable model for dimensionality reduction, IEEE Trans. Syst. Man. Cybern. Part B 41 (2) (2011) 425–434.
[25] X. Gao, C. Tian, Multi-view face recognition based on tensor subspace analysis and view manifold modeling, Neurocomputing 72 (16–18) (2009) 3742–3750.
[26] B. Xie, Y. Mu, D. Tao, K. Huang, m-SNE: multiview stochastic neighbor embedding, IEEE Trans. Syst. Man. Cybern. Part B 41 (4) (2011) 1088–1096.
[27] X. Wang, D. Tao, Z. Li, Entropy controlled Laplacian regularization for least square regression, Signal Process. 90 (6) (2010) 2043–2049.
[28] F. Samaria, A. Harter, Parameterisation of a stochastic model for human face identification, in: Proceedings of the IEEE Workshop, Appl. Comput. Vision (1994) 138–142.
[29] J. Tenenbaum, V. Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[30] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[31] T. Zhang, D. Tao, X. Li, J. Yang, Patch alignment for dimensionality reduction, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1299–1313.
[32] T. Zhang, D. Tao, J. Yang, Discriminative locality alignment, ECCV (2008) 725–738.
[33] X. He, D. Cai, J. Han, Learning a maximum margin subspace for image retrieval, IEEE Trans. Knowl. Data Eng. 20 (2) (2008) 189–201.
[34] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, Int. J. Comput. Vision 73 (2) (2001) 213–238.
[35] D.V. Jadhav, R.S. Holambe, Radon and discrete cosine transforms based feature extraction and dimensionality reduction approach for face recognition, Signal Process. 88 (10) (2008) 2604–2609.

**Dapeng Tao** received the B.S. degree in Electronics and Information Engineering from Northwestern Polytechnical University, Xi'an, China. He is currently a PhD candidate in Information and Communication Engineering at South China University of Technology, Guangzhou, China. His research interests include machine learning and computer vision.

**Lianwen Jin** received a B.S. degree from the University of Science and Technology of China and a Ph.D. degree from South China University of Technology in 1991 and 1996, respectively. He visited Motorola China Research Center in 2000 and the University of Hong Kong in 2002 and 2006 as a research fellow, respectively. He is now a professor at the School of Electronic and Information Engineering, South China University of Technology. He received the New Century Excellent Talent Program Award of China MOE in 2006. He has published more than 90 papers in the areas of handwritten Chinese character recognition, image processing and pattern recognition. His current research interests include character recognition, pattern analysis and recognition, image processing, machine learning and intelligent system. He is a member of the IEEE and IEEE Computer Society. He served as Program Committee member for a number of international conferences, including IWFHR '06, ICFHR '08, ICDAR '09, ICFHR '10, ICDAR2011 et.al.