Uncorrelated slow feature discriminant analysis using globality preserving projections for feature extraction

Xingjian Gu, Chuancai Liu, Sheng Wang, Cairong Zhao, Songsong Wu

Abstract

Slow Feature Discriminant Analysis (SFDA) is a supervised feature extraction method for classification inspired by biological mechanism. However, SFDA only considers the local geometrical structure information of data and ignores the global geometrical structure information. Furthermore, previous works have demonstrated that uncorrelated features of minimum redundancy are effective for classification. In this paper, a novel method called uncorrelated slow feature discriminant analysis using globality preserving projections (USFDA-GP) is proposed for feature extraction and recognition. In USFDA-GP, two kinds of global information are imposed to the objective function of conventional SFDA for respecting some more global geometric structures. We also provide an analytical solution by simple eigenvalue decomposition to the optimal model instead of previous iterative method. Experimental results on Extended YaleB, CMU PIE and LFW-a face databases demonstrate the effectiveness of our proposed method.

1. Introduction

Feature extraction is a fundamental and challenging issue in pattern recognition and machine learning. It has been widely used in many practical applications [1–3], such as face recognition, image analysis and so on. The goal of feature extraction is to seek low dimensional representations of high dimensional data such that the intrinsic structures of original high dimensional data are revealed. The most famous linear feature extraction methods, such as Principal Component Analysis (PCA) [4], Linear Discriminant Analysis (LDA) [5] and their extended methods [6,7], have been extensively used for face recognition.

Linear feature extraction methods may fail to discover the nonlinear structure of data. Recently, many geometrically motivated methods have been developed to discover nonlinear structure of high dimensional data. The representative nonlinear learning methods include Isomap [8], Locally Linear Embedding (LLE) [9], and Laplacian Eigenmap (LE) [10]. Isomap preserves pairwise geodesic distance of observations in embedding space. LLE focuses on local neighborhood of each data point and preserves the minimal linear reconstructing with neighborhood in the embedding space. LE is developed on Laplace Beltrami operator to preserve proximity relationship of data points.

However those manifold learning methods obtain low dimensional embedding without an explicit mapping, and they cannot extract feature beyond training samples. In order to overcome the problem, NPE [11] tries to find a linear subspace that preserves local structure based on the same principle of LLE. LPP [12] seeks a linear subspace to approximate nonlinear Laplacian Eigenmap.

In order to extract discriminant feature for classification, a lot of manifold learning based discriminant methods have been proposed to preserve the intrinsic geometry of the local neighborhoods and simultaneously reveal the discriminant structure of data [13–16]. The most prevalent approaches are Margin Fisher Analysis (MFA) [14], Discriminant Locality Preserving Projections (DLPD) [13], Local Discriminant Embedding (LDE), [15] and Local Discriminant Projection [16]. These approaches not only preserve the local geometrical structure, which represents the intra-class compactness, but also maximize the margin of inter-class to enhance the ability of classification in the reduced space.

Recently, more and more researches focus on applying the biological model to complex information tasks. Slow Feature Analysis (SFA), proposed by Wiskott and Sejnowski [17], extracted invariant from vectorial temporal signals based on temporal slowness principle. SFA has been applied for classification tasks in various ways. Franzius et al. [18] extracted the identity information of animated fish invariant to pose (including a rotation angle and the fish position) using SFA. Klampfl and Maass [19] introduced a particular Markov chain to generate a sequence of time
series used to train SFA for classification. Recently, more and more researchers have performed SFA to the applications of pattern recognition such as face recognition, human gesture recognition, human action recognition and other recognition tasks [20–25]. Zhang et al. [20] proposed the SFA framework to deal with the problem of human action recognition. SFA has a good performance on the data sets with temporal structure. However, in real applications, there are many discrete data sets that have no obvious temporal structure. In discrete scenario, it is necessary to generate time series before implementation of SFA. Huang et al. [23,25] utilized KNN criterion to construct time series and introduced Supervised Slow Feature Analysis (SSFA) for nonlinear dimensionality reduction. Gu et al. [21] proposed a new Supervised Slow Feature Analysis based on Consensus Matrix (SSFACM) to construct time series for face recognition. In [22], another variant of Supervised Slow Feature that seeks the Shortest Path of each class samples (SSFASP) was proposed to construct time series for dimensionality reduction. In order to get discriminant slow feature, Huang et al. [23] propose Slow Feature Discriminant Analysis (SFDA), which minimizes within-class temporal variation and maximize between-class temporal variation simultaneously for handwritten digit recognition. The same approach is also applied more recently to human gesture recognition by Koch et al. [24]. In summary, SFDA is also a local discriminant approach. SFDA encodes the discriminative information by maximizing the distance among nearby data points from different classes and preserves the intrinsic geometrical structure by minimizing the distance among nearby data from the same class. In the ideal case, nearby points from the same class will be mapped to a single point. Thus, SFDA only captures the local structure of data and ignores the global geometrical properties, resulting in unstable intrinsic structure representation.

In real world applications, the unknown structure of data is always complex. Thus, a single local geometrical structure may not be sufficient to represent the intrinsic geometric structure of data. A reasonable approach should be one that integrates both global and local structure into the objective function of feature extraction [26–29]. LDA [5] extracts discriminant feature based on global geometric structure of data. Thus, both of LapLDA [26] and Semi-supervised Discriminant Analysis (SDA) [27] represent the local geometry by LPP [12] and then integrate the local geometry into LDA. In the literature [28], the authors proposed Joint Global and Local structure Discriminant Analysis method (JGLDA), which used two quadratic functions characterize the geometric properties of similarity and diversity of data. Zhang et al. [29] proposed Complete Global–Local LDA (CGLDA) method to incorporate three kinds of local information: local similarity information, local intra-class pattern variation and local interclass pattern variation into LDA. All of the above mentioned methods represent the global structure based on the LDA framework. However, one problem often encountered in LDA based application is that Fisher criterion is not optimal for a c-class (c > 2) classification task. The reasons may contain two aspects. First, LDA overemphasizes the classes with larger distance in the original high dimensional space and causes large overlaps of neighboring classes in the low dimensional space [30,31]. Second, LDA assumes data covariances for all classes to be exactly identical [32], ignoring the diversity distribution of each class.

Furthermore, previous works [33–35] have demonstrated that statistically correlated features contain redundancy, which may distort the distribution of the feature and even dramatically degrade the performance. Recently, several uncorrelated discriminant methods have been developed. Jin et al. [33] proposed an uncorrelated linear discrimination analysis (ULDA) approach which maximizes Fisher criterion and simultaneously produces statistical uncorrelated features. To explore local information, Jing et al. [34] proposed a feature extraction approach named Local Uncorrelated Discriminant Transform (LUDT) for face recognition by constructing the local uncorrelated constraints and calculating the optimal discriminant vectors. However, the above mentioned uncorrelated methods are all implemented in an iterative way and it needs to take a long time to complete the iterative process.

In this paper, we propose a novel feature extraction method namely Uncorrelated Slow Feature Discriminant Analysis using Globality Preserving Projection (USFDA-GP), which integrates globality information into SFDA and extracts statically uncorrelated discriminant feature for classification. Some aspects of the proposed USFDA-GP method are worth highlighting.

1. This paper proposes a novel uncorrelated slow feature discriminant analysis method. The proposed method integrates globality information into traditional SFDA and removes the redundancy between features to enhance the discriminant ability.

2. USFDA-GP offers analytical solutions using standard eigenvalue decomposition and avoids the iterative process which is very time consuming. That is, USFDA-GP is more efficient in computation.

3. The features extracted by USFDA-GP, which integrates the global information of data into SFDA, are proven to have good discrimination ability. A series of experimental results show that USFDA-GP has many advantages on efficiency and accuracy in classification tasks.

The rest of the paper is organized as follows. In Section 2, we briefly review LDA and SFDA. In Section 3, we give the motivations of Uncorrelated Slow Feature Discriminant Analysis using Globality Preserving Projection and describe it in detail. In Section 4, experiments with face image databases are carried out to demonstrate the effectiveness of the proposed method. Finally, the conclusions are made in Section 5.

2. Related work

Given a sample set \( X = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{D \times n} \) and samples belong to one of c classes \( \{X_1, X_2, \ldots, X_c\} \). Let \( c \) denote the total number of classes and \( n_i \) denote the number of training samples in the ith class. Let \( \lambda_j \) denote the jth sample in the ith class, \( \bar{\lambda} \) denote the mean of all training samples, \( \bar{x}_i \) be the mean of the ith class.

2.1. LDA

LDA extracts discriminant feature for classification based on the principle of maximizing between-class scatter and minimizing within-class scatter simultaneously. The between-class and within-class scatter matrices can be evaluated as follows:

\[
S_b = \frac{1}{c} \sum_{i=1}^{c} (\bar{x}_i - \bar{\lambda}) (\bar{x}_i - \bar{\lambda})^T
\]

(1)

\[
S_w = \frac{1}{n} \sum_{i=1}^{n} (\lambda_i - \bar{x}_i) (\lambda_i - \bar{x}_i)^T
\]

(2)

The LDA based discriminant rule is defined as follows:

\[
W^* = \arg \max_W \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)}
\]

(3)

The solution to Eq. (3) can be solved by generalized eigenvalue problem \( S_b W = \lambda S_w W \) and optimal projections can be selected as eigenvectors \( w_1, w_2, \ldots, w_d \) corresponding to the first largest eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_d \).
2.2. Slow feature discriminant analysis

In this section, we present the detail of slow feature discriminant analysis (SFDA) [23] used for discrete data that does not have an obvious temporal structure. It firstly constructs within-class time series $t_w$ and between-class time series $t_b$ using neighboring information.

$$t_w = ([x^i_1, x^i_2]), i = 1, 2, \ldots, c, p \neq q, p q = 1, 2, \ldots, n_i$$ (4)

where $x^i_1$ and $x^i_2$ belong to ith class.

$$t_b = ([x^i_1, x^i_2]), \ i \neq j, \ i j = 1, 2, \ldots, c, p = 1, 2, \ldots, n_i, j = 1, 2, \ldots, n_j$$ (5)

where $x^i_1$ and $x^i_2$ belong to different classes.

Based on the set of time-series $t_w$ and $t_b$, the temporal variation $\Delta t_w$ and $\Delta t_b$ can be approximated by the time difference, where $\Delta t_w = ([x^i_1 - x^i_2]), (x^i_1, x^i_2) \in t_w$ and $\Delta t_b = ([x^i_1 - x^i_2]), (x^i_1, x^i_2) \in t_b$. The model of SFDA is as follows:

$$\min_{w, b} \max W^{T} T_w W + \sum_{i=1}^{c} \frac{n_i}{n} W^{T} W_{i} W$$

where $T_w = (1/n_i) \Delta t_w \Delta t_w^{T}$, $T_b = (1/n_b) \Delta t_b \Delta t_b^{T}$, $n_w$ and $n_b$ are the number of columns of $\Delta t_w$ and $\Delta t_b$ respectively. The solution to the optimization of Eq. (6) can be solved by eigenvalue problem $W^{T} W = \lambda T W$ and optimal projections can be selected as eigenvectors $w_1, w_2, \ldots, w_d$ corresponding to the first smallest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$.

3. Uncorrelated slow feature discriminant analysis using globality preserving projections

3.1. Motivation

The motivation for USFDA-GP arises from the following two considerable aspects. First, SFDA has been successfully performed to extract slow discriminant feature for classification, but it ignores the global geometric structure. Recent literatures [36,37] indicate that preserving the global structure in the low dimensional subspace is of great important for feature extraction. In this paper, we employ two kinds of models to preserve the globality geometric structure in the feature space. One is used to preserve the relationship of class centers in the low dimensional space. The other one is used to preserve each class covariance in the low dimensional feature space. Second, previous works have demonstrated that statically uncorrelated feature is favor for classification. However, the computing cost of the most uncorrelated methods is expensive, because most of them take an iterative manner to obtain the uncorrelated feature. We offer an analytical solutions using standard eigenvalue decomposition to avoid the iteration process.

![Fig. 1. Two dimensional points: the points with different colors or shapes are belonging to different classes (a), (b) and (c) respectively. Points of class (a) reside on a compact region. Points of class (b) and class (c) reside on sparse regions. $d(a, c), d(a, b), d(b, c)$ are the distances between class centers.](image)

Based on the aforementioned analysis, we propose USFDA-GP to obtain statistically uncorrelated slow discriminant feature, which not only considers the local structure of manifold but also the global geometric structure.

3.2. Globality preserving projections

First, we use the center of each class to capture the global structure of data. Inspired by LDA [5], maximizing the margin between class centers is helpful to extract discriminant feature. However, it will result that the nearby classes are likely overlapped in the feature space. To obtain well performance of classification, the pair of classes with smaller distances should be given more weight and two classes which are already well-separated can have less weight in the objective function. We define a new weighted objective function as follows:

$$\max_{W} \text{trace} \left( W^{T} \left( \sum_{i=1}^{c} \frac{n_i}{n} W_{i} W_{i}^{T} \right) W \right)$$ (7)

where $w(d_{ij}) = (d_{ij})^{-1}$ is the weighting function that depends on the distance $d_{ij}$ between the ith class and the jth class, $d_{ij} = ||x_i - x_j||^2$ and hyper-parameter $t > 0$. In this objective function (7), the pair of classes with smaller distances contribute more than the pair of classes with larger distances. Parameter $t$ controls the quantity of contribution from the pair of classes with smaller distances. The larger the $t$ is, the weighting function with small distance will have a larger weight value. From Fig. 1, it is obvious that $d(a, c) > d(a, b) > d(b, c)$, class (a) and class (c) are already well-separated while class (b) and class (c) are nearby. Thus, emphasizing the margin between class (b) and class (c) will receive a better performance of classification. It is consistent with our common sense. Therefore, the definition of objective (7) is reasonable.

Second, let us consider preserving class covariance in the low dimensional feature space. From Fig. 1, it is easy to see that the distribution of each class is different, i.e. the points of class (a) reside on a compact region and points of class (b) and class (c) reside on sparse regions. They do not share an identical distribution. Since covariance is a good tool to describe data distribution. Since covariance is a good tool to describe data distribution, preserving class covariance can preserve the distribution of data in the low dimensional feature space. Based on minimizing the error of the reconstruction using Two-dimensional Principal Component Analysis (2DPCA)[6], we have the following model:

$$\min_{W} \frac{n}{\sum_{i=1}^{c} n_i} ||\left( S_w - S_w \right) - W W^{T} \left( S_w - S_w \right)||^2$$ (8)

where $S_w$ is the covariance matrix of the ith class, $S_w = \frac{1}{n} \sum_{j=1}^{n_i} x_i$ is the mean of covariance matrix and $W \in R^{p \times d}$ is the transform matrix.

It is easy to see that

$$\frac{n}{\sum_{i=1}^{c} n_i} ||\left( S_w - S_w \right) - W W^{T} \left( S_w - S_w \right)||^2$$

$$\begin{align*}
= & \text{trace} \left( \left( S_w - S_w \right) - W W^{T} \left( S_w - S_w \right) \right)^{T} \left( S_w - S_w \right) \\
= & \text{trace} \left( S_w - S_w \right)^{T} \left( S_w - S_w \right) - 2 \times \text{trace} \left( S_w - S_w \right)^{T} W W^{T} \left( S_w - S_w \right) \\
+ & \text{trace} \left( S_w - S_w \right)^{T} W W^{T} \left( S_w - S_w \right) \\
= & \text{trace} \left( S_w - S_w \right)^{T} \left( S_w - S_w \right) - \text{trace} \left( S_w - S_w \right)^{T} W W^{T} \left( S_w - S_w \right) \\
= & \text{trace} \left( S_w - S_w \right)^{T} \left( S_w - S_w \right) \text{trace} \left( W^{T} \left( S_w - S_w \right) \right) \\
= & \text{trace} \left( S_w - S_w \right)^{T} \left( S_w - S_w \right) \text{trace} \left( W^{T} \left( S_w - S_w \right) \right)
\end{align*}$$
Thus, the objective (8) can be written as follows:

$$\max_{\mathbf{W}} \text{trace} \left( \mathbf{W}^T \sum_{i=1}^{C} (\mathbf{S}_w - \mathbf{S}_w) \mathbf{W} \right)$$

(9)

### 3.3. Uncorrelated constraint

In this subsection, we consider the statistically uncorrelated constraint. It has been found in the literatures [33,38] that the extracted discriminant feature are actually statistically uncorrelated over total data, that is for any two different low dimensional embedding feature \(y_i\) and \(y_j\) (\(i \neq j\)) of the extracted feature \(Y = \mathbf{W}^T \mathbf{X}\) are statistically uncorrelated with the following constraints:

$$w_i^T S_w w_j = 0, \quad i \neq j$$

(10)

We normalize \(w_i\) to satisfy

$$w_i^T S_w w_i = 1$$

(11)

where \(S_w = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T\) is the total data covariance and \(\bar{x}\) denotes the mean of all training samples.

Thus, the statistically uncorrelated constraint can be summarized as

$$W^T S_w W = I$$

(12)

where \(I\) is identity matrix.

### 3.4. Objective and solution

For the convenience of writing, we denote \(S_b = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \left( n_i/n \right) w(d_{ij})(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T\) and \(P_{S_w} = \sum_{i=1}^{C} (S_w - \mathbf{S}_w)^2\). From the above analysis, the objective of USFDA-GP can be written as follows:

$$\min_{\mathbf{W}} \text{trace} \left( \mathbf{W}^T (T_w - \alpha(T_b + S_b + P_{S_w})) \mathbf{W} \right)$$

s.t. \(W^T S_t W = I\)

(13)

where \(T_w\) is within-class temporal variation, \(T_b\) is between-class temporal variation and \(\alpha\) is the tradeoff parameter.

**Lemma 1.** The total scatter matrix \(S_t \in R^{D \times D}\) can be rewritten as \(S_t = HH^T\), where \(H \in R^{D \times D}\).

**Lemma 2.** \((H^T)^{-1} = (H^{-1})^T\)

**Theorem 1.** The solution of objective function (13) can be obtained by standard eigenvalue decomposition. The optimal solution of objective function 13 is \(W = \mathbf{H} V\), where \(V = [v_1, v_2, ..., v_d]\) is the eigenvectors of \(H^{-1}(T_w - \alpha(T_b + S_b + P_{S_w}))\) corresponding to the first smallest eigenvalues \(\lambda_1, \lambda_2, ..., \lambda_d\).

### 3.5. Algorithm of USFDA-GP

As analyzed in previous subsections, the algorithmic procedure of USFDA-GP is formally summarized as follows:

**Step 1:** Construct matrices: \(T_w, T_b, S_b, S_t\) and \(P_{S_w}\).

**Step 2:** Perform Singular Value Decomposition to \(S_t = P \Lambda P^T\) and compute \(H = P \Lambda^{1/2}\).

**Step 3:** Solve the eigenvalue problem \(H^{-1}(T_w - \alpha(T_b + S_b + P_{S_w}))\) \((H^{-1}) V = \lambda V\) and select eigenvector \(V = [v_1, v_2, ..., v_d]\) corresponding to the first smallest eigenvalues \(\lambda_1, \lambda_2, ..., \lambda_d\).

---

**Fig. 2.** Sample images in Extended YaleB database.

**Fig. 3.** Sample images in PIE database.
Step 4: Obtain the optimal transformation matrix $W = P A^{-1/2} V$ for a new sample $x$, its low-dimensional feature representation $y = W^T x$.

4. Experiments

In this section, we compare the classification performance of the proposed USFDA-GP with some representative feature extraction methods including LDA [5], DLPP [13], MFA [14], MMC [39], SFDA [23], and ULDA [33] on Extended YaleB, CMU PIE and LFW-a face image databases. In order to evaluate the effectiveness of preserving globality structure of our method, we also compare with the Uncorrelated Slow Feature Discriminant Analysis (USFDA). To avoid the small sample size (SSS) problem, PCA is first used as preprocessing step, where we keep nearly 98% data.

### 4.1. Databases

The Extended YaleB Face Database contains 16,128 images under 9 poses and 64 illumination conditions. In our experiment, we select a subset contains 2431 images of 38 individuals. Before conduct our experiment, we crop the face portion of the image into the resolution of $32 \times 32$. Fig. 2 shows some sample images.

The CMU PIE Face Database contains 41,368 images from 68 individuals. These images of each individual were taken under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our experiment, we select a subset that contains 11,554 images of 68 individuals. Before implement our experiment, we crop the face portion of the image into the resolution of $32 \times 32$. Some sample images are shown in Fig. 3.

The LFW-a database [44] is an automatically aligned gray scale version of the LFW database [45] which is a database that aims at studying the problem of the unconstrained face recognition. The database is considered as one of the most challenging database since

---

**Table 1**

<table>
<thead>
<tr>
<th>Method</th>
<th>8 Train.</th>
<th>10 Train.</th>
<th>12 Train.</th>
<th>15 Train.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMC</td>
<td>74.47 ± 2.04</td>
<td>78.32 ± 2.43</td>
<td>81.92 ± 2.03</td>
<td>84.46 ± 1.34</td>
</tr>
<tr>
<td>MFA</td>
<td>71.61 ± 2.21</td>
<td>76.52 ± 1.82</td>
<td>79.89 ± 1.45</td>
<td>82.93 ± 1.71</td>
</tr>
<tr>
<td>DLPP</td>
<td>74.83 ± 1.53</td>
<td>78.63 ± 1.74</td>
<td>82.30 ± 1.63</td>
<td>85.40 ± 1.28</td>
</tr>
<tr>
<td>LDA</td>
<td>74.93 ± 1.70</td>
<td>78.62 ± 1.99</td>
<td>82.27 ± 1.66</td>
<td>85.29 ± 1.29</td>
</tr>
<tr>
<td>ULDA</td>
<td>76.27 ± 2.03</td>
<td>80.18 ± 2.20</td>
<td>84.57 ± 1.55</td>
<td>86.99 ± 1.44</td>
</tr>
<tr>
<td>SFDA</td>
<td>74.88 ± 1.68</td>
<td>79.27 ± 2.05</td>
<td>83.02 ± 2.05</td>
<td>86.03 ± 1.18</td>
</tr>
<tr>
<td>USFDA</td>
<td>75.83 ± 1.96</td>
<td>80.11 ± 1.87</td>
<td>84.12 ± 1.55</td>
<td>86.95 ± 1.10</td>
</tr>
<tr>
<td>USFDA-GP</td>
<td>78.42 ± 1.70</td>
<td>81.94 ± 2.04</td>
<td>85.44 ± 2.08</td>
<td>88.05 ± 1.19</td>
</tr>
</tbody>
</table>

**Table 2**

<table>
<thead>
<tr>
<th>Method</th>
<th>8 Train.</th>
<th>10 Train.</th>
<th>12 Train.</th>
<th>15 Train.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMC</td>
<td>71.62 ± 2.55</td>
<td>76.97 ± 2.6</td>
<td>79.83 ± 2.26</td>
<td>83.70 ± 2.51</td>
</tr>
<tr>
<td>MFA</td>
<td>64.23 ± 2.34</td>
<td>71.80 ± 1.7</td>
<td>75.92 ± 2.82</td>
<td>81.47 ± 1.90</td>
</tr>
<tr>
<td>DLPP</td>
<td>75.15 ± 1.98</td>
<td>80.33 ± 2.4</td>
<td>82.60 ± 2.96</td>
<td>86.22 ± 2.14</td>
</tr>
<tr>
<td>LDA</td>
<td>75.42 ± 2.17</td>
<td>80.43 ± 2.17</td>
<td>82.50 ± 2.56</td>
<td>86.52 ± 1.74</td>
</tr>
<tr>
<td>ULDA</td>
<td>77.55 ± 2.44</td>
<td>81.28 ± 2.28</td>
<td>83.45 ± 2.75</td>
<td>86.83 ± 2.46</td>
</tr>
<tr>
<td>SFDA</td>
<td>74.18 ± 2.07</td>
<td>78.92 ± 2.38</td>
<td>81.93 ± 2.18</td>
<td>85.53 ± 2.14</td>
</tr>
<tr>
<td>USFDA</td>
<td>77.33 ± 2.23</td>
<td>81.20 ± 2.29</td>
<td>83.65 ± 2.46</td>
<td>87.68 ± 1.58</td>
</tr>
<tr>
<td>USFDA-GP</td>
<td>81.37 ± 1.95</td>
<td>83.47 ± 2.11</td>
<td>86.20 ± 1.73</td>
<td>88.70 ± 1.68</td>
</tr>
</tbody>
</table>

---

Fig. 4. Sample images in LFW-a database.
it contains 13,233 images from 5749 different people, with great variations in terms of lighting, pose, age, and even image quality. We crop the face images as in [46] and gather the subjects including no less than 40 samples to get a sub-database with 19 subjects from LFW-a. Fig. 4 shows some sample images in the LFW-a database.

4.2. Experiment on the YaleB and PIE database

In this experiment, we evaluate the robustness of each method to the changes of poses, illumination conditions and expressions on the YaleB and PIE face databases. Specially, we randomly select \( l = (8, 10, 12, 15) \) samples from each individual for training, and the rest of samples are used for testing. For each given \( l \), we repeat each experiment 20 times and calculate the average recognition accuracy. The maximal average recognition accuracy obtained by different dimensionality reduction methods as well as standard deviations and the corresponding dimensionality of reduced sub-space are given in Tables 1 and 2.

From the results shown in Tables 1 and 2, all methods are improved significantly when the number of training sample increases. The reason is that a large set of training data can sample the underlying distribution more accurately than a smaller set. Figs. 5 and 6 demonstrate the recognition accuracy of different methods over the variance of the dimensionality of subspaces. In addition, more can be found from Figs. 5 and 6. First, the proposed method USFDA-GP outperforms other methods. The good performance of USFDA-GP demonstrates that USFDA-GP is more effective than other methods in feature extraction. Second, increasing dimension will improve the recognition rate until the dimension reaches some optimal value. In particular, the recognition rate curve of our method always arrives its best value much earlier (when a small dimension is used) than other compared methods. It therefore seems clearly that USFDA-GP is more effective than other compared methods in accounting for discriminant information. Third, the recognition of USFDA-GP is superior to that of the local discriminant methods, such as DLPP, MFA, SFDA and USFDA. The main reason may be that those local discriminant methods only preserve the local geometric structure and cannot preserve distribution of each class in the low dimensional feature space. Fourth, USFDA-GP is also superior to MMC, LDA and ULDA. This is probably because MMC, LDA and ULDA do not well encode the discriminative information embedded in nearby data from different class and cannot preserve the local structure in the low dimensional feature space. Moreover, LDA neglects the diversity of margin between pairs of class centers. As a result, there is a large overlap among those nearby classes, leading to a low classification rate. Hence, our method has better discriminating power than the compared methods in extracting and representing facial features for face recognition against the variation of lighting, facial expressions and pose.
Fig. 6. Average recognition rate versus reduced dimensionality on PIE database with $l = \{8, 10, 12, 15\}$. (a) $l = 8$, (b) $l = 10$, (c) $l = 12$, and (d) $l = 15$.

Fig. 7. Sample images from LFW-a according to six different face resolutions. (a) $10 \times 10$, (b) $20 \times 20$, (c) $30 \times 30$, (d) $40 \times 40$, (e) $50 \times 50$, and (f) $60 \times 60$.

Table 3

<table>
<thead>
<tr>
<th>Method</th>
<th>10 x 10</th>
<th>20 x 20</th>
<th>30 x 30</th>
<th>40 x 40</th>
<th>50 x 50</th>
<th>60 x 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMC</td>
<td>45.13 ± 2.78 (17)</td>
<td>59.40 ± 2.30 (33)</td>
<td>63.17 ± 2.35 (26)</td>
<td>63.97 ± 2.29 (18)</td>
<td>64.54 ± 2.29 (27)</td>
<td>64.55 ± 2.47 (21)</td>
</tr>
<tr>
<td>MFA</td>
<td>48.16 ± 2.33 (24)</td>
<td>60.81 ± 2.67 (37)</td>
<td>62.16 ± 2.68 (30)</td>
<td>62.55 ± 2.42 (33)</td>
<td>62.04 ± 2.38 (34)</td>
<td>61.95 ± 2.86 (36)</td>
</tr>
<tr>
<td>DLPP</td>
<td>52.21 ± 2.74 (16)</td>
<td>69.59 ± 2.43 (18)</td>
<td>73.17 ± 2.15 (18)</td>
<td>73.93 ± 1.75 (18)</td>
<td>74.26 ± 1.83 (18)</td>
<td>74.32 ± 2.10 (18)</td>
</tr>
<tr>
<td>LDA</td>
<td>52.07 ± 2.50 (18)</td>
<td>69.47 ± 2.59 (18)</td>
<td>73.08 ± 2.29 (18)</td>
<td>74.00 ± 1.98 (18)</td>
<td>74.41 ± 1.86 (18)</td>
<td>74.31 ± 2.11 (18)</td>
</tr>
<tr>
<td>SFDA</td>
<td>54.97 ± 2.02 (23)</td>
<td>71.69 ± 2.21 (28)</td>
<td>74.68 ± 2.17 (28)</td>
<td>74.89 ± 2.16 (30)</td>
<td>75.19 ± 2.36 (27)</td>
<td>75.17 ± 2.38 (27)</td>
</tr>
<tr>
<td>USFDA</td>
<td>45.37 ± 2.49 (39)</td>
<td>65.19 ± 2.33 (42)</td>
<td>70.23 ± 1.79 (51)</td>
<td>70.93 ± 1.48 (50)</td>
<td>71.67 ± 1.43 (48)</td>
<td>71.75 ± 1.35 (50)</td>
</tr>
<tr>
<td>USFDA-GP</td>
<td>52.12 ± 2.18 (44)</td>
<td>68.40 ± 2.09 (45)</td>
<td>72.78 ± 1.84 (65)</td>
<td>73.86 ± 1.42 (40)</td>
<td>74.56 ± 1.39 (40)</td>
<td>76.61 ± 1.42 (37)</td>
</tr>
<tr>
<td>USFDA-GP</td>
<td>57.24 ± 3.06 (24)</td>
<td>72.88 ± 2.24 (20)</td>
<td>76.59 ± 2.71 (20)</td>
<td>77.14 ± 2.79 (23)</td>
<td>77.56 ± 1.50 (17)</td>
<td>77.41 ± 1.92 (17)</td>
</tr>
</tbody>
</table>
4.3. Experiment on the LFW-a database

The aim of this experiment is to access the performance of each method to the variation of lighting, pose, age and even image quality using different resolution images. We design the comparative experiment on six different face resolutions, i.e. $10 \times 10$, $20 \times 20$, $30 \times 30$, $40 \times 40$, $50 \times 50$ and $60 \times 60$. To acquire face images with varying resolutions, we carry out resizing over the LFW-a dataset. Fig. 7 shows the examples of face images with six different resolutions used in our experiment. As in the previous experiments, 15 images of each individual are selected for training, and the rest images are used for testing. The best average recognition accuracy corresponding as well as the standard deviations and the optimal dimensionality of reduced subspace over 20 runs of tests are shown in Table 3. Fig. 8 illustrates
the curves of recognition accuracy vs. the variations of the dimensions using six different resolution face images.

From the results, it can be seen that the recognition accuracies of the eight feature extraction methods are relatively low on the LFW-a database. It is because that the light and expression of each person has great variation, the face images often be misalignment and a large number of images on LFW-a are low quantity. However, compared with MMC, DLPP, MFA, LDA, ULDA, SFDA and USFDA, our method has an apparent advantage. Fig. 8 shows that our proposed method also consistently outperforms over the other methods. The good performance of USFDA-GP also demonstrates that USFDA-GP is more effective than other methods in feature extraction. From Table 3 and Fig. 8, we also find that the uncorrected feature extracted methods have a better performance for classification, i.e. ULDA outperforms LDA and USFDA outperforms SFDA. It demonstrates that removing the correlation between features can enhance the discriminant ability.

Also, we discuss the computational cost of our proposed USFDA-GP in comparison to MMC, DLPP, MFA, LDA, ULDA, SFDA and USFDA. To extract a set of uncorrelated feature, previous method takes an iterative manner, which is time consuming. As mentioned in Eq. (A.6), USFDA-GP can take a standard eigenvalue decomposition to avoid the iteration process to extract uncorrelated feature. We select \( l = 8 \) samples of each individual to calculate the training cost of each method. Each experiment is randomly repeatedly performed 20 times. We list the average computational time of all methods using 20 \( \times \) 20, 40 \( \times \) 40 and 60 \( \times \) 60 face images in Table 4. From the result, we can see that the higher resolution image will cost more time, and our method USFDA-GP is less time consuming than other uncorrelated feature method, such as ULDA and USFDA.

### Table 4
The average training time (second) of MMC, DLPP, MFA, LDA, ULDA, SFDA, USFDA and USFDA-GP across 20 runs on LFW-a databases.

<table>
<thead>
<tr>
<th>Method</th>
<th>MMC</th>
<th>DLPP</th>
<th>MFA</th>
<th>LDA</th>
<th>ULDA</th>
<th>SFDA</th>
<th>USFDA</th>
<th>USFDA-GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 ( \times ) 20</td>
<td>0.073</td>
<td>0.178</td>
<td>0.420</td>
<td>0.050</td>
<td>0.592</td>
<td>2.199</td>
<td>0.669</td>
<td></td>
</tr>
<tr>
<td>40 ( \times ) 40</td>
<td>0.151</td>
<td>0.252</td>
<td>2.774</td>
<td>0.110</td>
<td>1.762</td>
<td>0.759</td>
<td>3.775</td>
<td>0.805</td>
</tr>
<tr>
<td>60 ( \times ) 60</td>
<td>0.160</td>
<td>0.432</td>
<td>4.24</td>
<td>1.927</td>
<td>0.910</td>
<td>4.482</td>
<td>1.003</td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 9. Maximal average recognition accuracy of USFDA-GP with different parameter \( \alpha \) using \( l = 8, 10, 12, 15 \) training samples on YaleB database.](image1)

![Fig. 10. Maximal average recognition accuracy of USFDA-GP with different parameter \( \alpha \) using \( l = 8, 10, 12, 15 \) training samples on PIE database.](image2)
4.4. Influence of parameter $\alpha$ and $t$

In this subsection, we first study the parameter $\alpha$ impact to the performance of USFDA-GP on YaleB, PIE and LFW-a face databases. The value of parameter $\alpha$ is set to be 0.0001, 0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500 and 1000. We randomly select $l$ ($=8, 10, 12, 15$) images from each individual for training, while the remaining images are used for testing on YaleB and PIE face databases.

![Fig. 11](image1.png) Maximal average recognition accuracy of USFDA-GP with different parameter $\alpha$ using different resolution face images on LFW-a database.

![Fig. 12](image2.png) Maximal average recognition accuracy versus parameter $t$ on (a) YaleB, (b) PIE, and (c) LFW-a face databases.
In this paper, we have developed a new feature extraction method called uncorrelated slow feature discriminant analysis using globality preserve projections (USFDA-GP). Essentially, the proposed method can be viewed as an extension of SFDA. USFDA-GP integrates global geometrical structure into the objective function of SFDA and extracts uncorrected discriminant feature for face recognition. To be specific, we maximize the weighted margin of class centers and preserving the covariance of each class to preserve the global structure in the low dimensional feature space. Moreover, USFDA-GP also provides an analytical solutions by simple eigenvalue decomposition, avoiding the iterative process as in ULDA [33], to extract uncorrected feature. Experimental results on three image databases (Extended YaleB, CMU PIE and LFW-a) demonstrate the effectiveness of the proposed method.

An intriguing question for future work is whether this framework can deal with some individuals who are not in the training set. The success of incremental learning has been noticed in the work [47] and minimization of L∞ norm is a powerful method for unknown samples detection [48]. We believe that the full potential of integration of incremental learning and unknown samples detection has not been well explored. From a practical standpoint, it would be useful to extend our algorithm to an online learning framework that can deal with the unknown persons in the process of face recognition.

Acknowledgments

This work is supported by the National Natural Science Fund of China (Grant nos. 61373062, 61373063, 61203247), the Project of Ministry of Industry, Information Technology of PRC (Grant no. E0310/1112/02-1) and Fundamental Research Funds for the Central Universities (Grant no. 2015B03114).

Appendix A. Appendix

Lemma 3. The total scatter matrix $S_t \in \mathbb{R}^{D \times D}$ can be rewritten as:

$$S_t = HH^T, \text{where } H \in \mathbb{R}^{D \times D}$$

Proof. Given $S_t$ is a semi-positive definite symmetric matrix, the singular value decomposition of $S_t$ can be written in the form of $S_t = P \Lambda P^T$, where $P \in \mathbb{R}^{D \times r}$ is a set of orthogonal basis, $\Lambda \in \mathbb{R}^{r \times r}$ is a diagonal matrix and $r$ is the rank of $S_t$. We denote $H = PA^{1/2}$, where $A = \Lambda^{1/2}\Lambda^{1/2}$, thus $S_t = HH^T$.

Lemma 4. $(H^T)^{-1} = (H^{-1})^T$

Proof. According to the property of singular value decomposition, we have the following results $PP^T = I$, $A$ is a diagonal matrix and $A = \Lambda^{1/2}\Lambda^{1/2}$. Thus, we can easily get

$$(H^{-1} = (PA^{1/2})^{-1} = (A^{-1/2}P^{-1})^T = PA^{-1/2}$$

$$(H^T)^{-1} = (A^{1/2}P^T)^T$$

Thus we get $(H^T)^{-1} = (H^{-1})^T$.

Theorem 2. The solution of objective function 13 can be obtained by standard eigenvalue decomposition. The optimal solution of objective function 13 is $W = (H^T)^{-1}V$, where $V = [v_1, v_2, \ldots, v_d]$ is the eigenvectors of $H^{-1}(T_w - \alpha(T_b + S_b + P_{bw}))(H^{-1})^T$, $v = \lambda v$ corresponding to the first smallest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$.

Proof. Because of $S_t = HH^T$, Eq. (12) can be rewritten as $V^T V = I$, where $V = H^TW$. According to Lemma 1 and Lemma 2, it is easy to get the following equations:

$$W = (H^T)^{-1} = (H^{-1})^T V$$

$$W^T = ((H^{-1})^T)^T = V^T (H^{-1})^T = V^T H^{-1}$$

Substituting Eqs. (A.4) and (A.5) into Eq. (13), the objective can be rewritten as

$$\min_{V} \langle V^T H^{-1}(T_w - \alpha(T_b + S_b + P_{bw}))(H^{-1})^T V \rangle$$

s.t. $V^T V = I$.

Since $H^{-1}(T_w - \alpha(T_b + S_b + P_{bw}))(H^{-1})^T$ is a symmetric matrix, the objective A.6 can be reduced to an eigenvalue problem $H^{-1}(T_w - \alpha(T_b + S_b + P_{bw}))(H^{-1}) V = \lambda V$ and we select eigenvector $V = [v_1, v_2, \ldots, v_d]$ corresponding to the first smallest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$. The final optimal projection matrix is $W = (H^T)^{-1} V$.

References

J. Chen, J. Ye, Q. Li, Integrating global and local structures: a least squares approach,

X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in:

X. He, P. Niyogi, Locality preserving projections, in: Neural Information

W. Yu, X. Teng, C. Liu, Face recognition using discriminant locality preserving

S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and

T. Zeng, Y. Shen, H. Wei, T. Liu, Local discriminant embedding and its

J. Hu, W. Deng, J. Guo, W. Xu, Learning a locality discriminating projection for

L. Wiskott, T.J. Sejnowski, Slow feature analysis: unsupervised learning of

M. Franzus, N. Willibert, L. Wiskott, Invariant Object Recognition with Slow
Feature Analysis, 2008.

S. Klampfl, W. Maass, Replacing supervised classification by Slow Feature


X. Gu, C. Liu, S. Wang, Supervised slow feature analysis for face recognition,

X. Gu, C. Liu, Z. Yang, Dimensionality reduction based on supervised slow

Y. Huang, J. Zhao, M. Tian, Q. Zou, S. Luo, Slow feature discriminant analysis
and its application on handwritten digit recognition, in: International Symp-
posium on Neural Networks, 2005, pp. 1294–297.

P. Koch, W. Konen, K. Hein, Gesture recognition on few training data using
slow feature analysis and parametric bootstrap, in: International Symposium
on Neural Networks, 2010, pp. 1–8.

Y. Huang, J. Zhao, Y. Liu, S. Luo, Q. Zou, M. Tian, Nonlinear dimensionality

J. Chen, J. Ye, Q. Li, Integrating global and local structures: a least squares
framework for dimensionality reduction([C]), IEEE Conference on Computer

D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: International

Q. Gao, J. Liu, H. Zhang, X. Gao, K. Li, Joint global and local discriminant

D. Zhang, J. He, Y. Zhao, Z. Luo, M. Du, Global plus local: a complete framework

J. Peng, R.P.W. Duin, R. Haeb-umhoch, Multiclass linear dimension reduction

Y. Liang, C. Li, W. Gong, Y. Pan, Uncorrelated linear discriminant analysis

D. Kong, C. Ding, Pairwise-covariance linear discriminant analysis, in: Twenty-
Eighth AAAI Conference on Artificial Intelligence, 2014.

Z. Jin, Y. Yang, Z. MinTang, Z. shanHu, A theorem on the uncorrelated optimal

X. Jing, S. Li, D. Zhang, J. Yang, Face recognition based on local uncorrelated
and weighted global uncorrelated discriminant transforms, in: IEEE Interna-

constrained by locally statistical uncorrelation for face recognition, Neural

Z. Lai, Z. Jin, J. Yang, Global sparse representation projections for feature
extraction and classification, in: Chinese Conference on Pattern Recognition,
2009.

Y.H. Yuan, Q.S. Sun, Multiset canonical correlations with globality preserving
projections with applications to feature extraction and recognition([J]), IEEE
Transactions on Neural Networks and Learning Systems 25 (6) (2014)
1131–1146.

J. Ye, T. Li, T. Xiong, R. Banard, Using uncorrelated discriminant analysis for

X. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum

G-F. Lu, Z. Lin, Z. Jin, Face recognition using discriminant locality preserving
projections based on maximum margin criterion, Pattern Recognit. 43 (10)

L. Zhang, L. Qiao, S. Chen, Graph-optimized locality preserving projections,

H. Tang, Y. Sun, D. Xu, X. Tang, T. Huang, Trace ratio vs. ratio trace
for dimensionality reduction, in: IEEE Conference on Computer Vision

H. Wang, F. Nie, H. Huang, Globally and locally consistent unsupervised

L. Wolf, T. Hassner, Y. Taigman, Effective unconstrained face recognition by

G.B. Huang, M. Ramesh, T. Berg, et al. Labeled faces in the wild: A database for
studying face recognition in unconstrained environments[ R]. Technical Report
07-49, University of Massachusetts, Amherst, 2007.

Z. Lei, M. Pietikainen, S.Z. Li, Learning discriminant face descriptor, IEEE Trans.

T.-K. Kim, K.-Y.Y. Wong, B. Stenger, J. Kittler, R. Cipolla, Incremental linear
discriminant analysis using sufficient spanning set approximations, in: IEEE

F. Shen, C. Shen, R. Hill, Andrew Hengel, Z. Tang, Fast approximate Lc minimiza-

Xingjian Gu is now working for Ph.D. degree at the School of Computer Science and Engineering of Nanjing University of Science and Technology. He received his B. S. degree in the College of Math and Physics at Nanjing University of Information Science and Technology in 2007. His research interests mainly focus on Pattern Recognition and Computer Vision.

Chuancai Liu is a Full Professor in the School of Computer Science and Engineering of Nanjing University of Science and Technology, China. He obtained his Ph.D. degree from the China Ship Research and Development Academy in 1997. His research interests include AI, Pattern Recognition and Computer Vision. He has published about 50 papers in International/National Journals.

Sheng Wang received B.S. degree in automation from Henan University, China, in 2008. He obtained his M.S. degree in Control Theory and Control Engineering from the same university. Currently, he is a Ph.D. student at Nanjing University of Science and Technology. His research interests include Image Processing, Pattern Recognition and Machine Learning.

Cairong Zhao is currently an assistant professor at Tongji University. He received the Ph.D. degree from Nanjing University of Science and Technology, M.S. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, and B.S. degree from Jilin University, in 2011, 2006 and 2003, respectively. His research interests include Face Recognition, Building Recognition and Vision Attention.

Songsoo Wu received the B.S. degree in Information and Computing Sciences from Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China, in 2004, the M.S. degree in Measuring and Testing Technologies and Instruments from Nanjing Forestry University (NJFU), Nanjing, China, in 2007, and the Ph. D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2012. He is currently a lecturer of Nanjing University of Posts and Telecommunications. His research interests include pattern recognition, machine learning, image processing and analysis, data compression and retrieval.