



Local uncorrelated local discriminant embedding for face recognition*

Xiao-hu MA^{†1,2}, Meng YANG¹, Zhao ZHANG¹

(¹School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

(²State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China)

E-mail: xhma@suda.edu.cn; eyangmeng@163.com; cszzhang@suda.edu.cn

Received Aug. 7, 2015; Revision accepted Dec. 2, 2015; Crosschecked Jan. 20, 2016

Abstract: The feature extraction algorithm plays an important role in face recognition. However, the extracted features also have overlapping discriminant information. A property of the statistical uncorrelated criterion is that it eliminates the redundancy among the extracted discriminant features, while many algorithms generally ignore this property. In this paper, we introduce a novel feature extraction method called local uncorrelated local discriminant embedding (LULDE). The proposed approach can be seen as an extension of a local discriminant embedding (LDE) framework in three ways. First, a new local statistical uncorrelated criterion is proposed, which effectively captures the local information of interclass and intraclass. Second, we reconstruct the affinity matrices of an intrinsic graph and a penalty graph, which are mentioned in LDE to enhance the discriminant property. Finally, it overcomes the small-sample-size problem without using principal component analysis to preprocess the original data, which avoids losing some discriminant information. Experimental results on Yale, ORL, Extended Yale B, and FERET databases demonstrate that LULDE outperforms LDE and other representative uncorrelated feature extraction methods.

Key words: Feature extraction, Local discriminant embedding, Local uncorrelated criterion, Face recognition

<http://dx.doi.org/10.1631/FITEE.1500255>

CLC number: TP391.4

1 Introduction

Face recognition has received increasing attention in computer vision and pattern recognition because of its special advantages. However, it is a challenge to analyze the high dimensionality of input data. To resolve this problem, numerous dimensionality reduction methods have been proposed in recent years. They aim to find a low-dimensional subspace of high-dimensional data and project the original data on it. The most popular dimensional-

ity reduction techniques may be principal component analysis (PCA) (Turk and Pentland, 1991) and linear discriminant analysis (LDA) (Belhumeur *et al.*, 1997). PCA aims to calculate the project vectors of the low-dimensional subspace, in which the covariance matrix of the training set is maximized. Unlike PCA which is unsupervised, LDA is a supervised feature extraction method, which tends to find a linear transformation that maximizes the interclass scatter and minimizes the intraclass scatter simultaneously. It is generally believed that LDA outperforms PCA.

Even though PCA and LDA preserve the global linear Euclidean structure of the original data, they fail to discover the underlying low-dimensional manifold structure. Many manifold learning algorithms have been proposed to resolve this issue. The theoretical foundation of manifold learning is based on

[†] Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61402310), the Natural Science Foundation of Jiangsu Province, China (No. BK20141195), and the State Key Laboratory for Novel Software Technology Foundation of Nanjing University, China (No. KFKT2014B11)

© ORCID: Xiao-hu MA, <http://orcid.org/0000-0002-2384-3137>
 © Zhejiang University and Springer-Verlag Berlin Heidelberg 2016

the assumption that the high-dimensional data lies on an intrinsic low-dimensional submanifold. The most representative manifold learning algorithms contain local linear embedding (LLE) (Roweis and Saul, 2000), Isomap (Tenenbaum *et al.*, 2000), and Laplacian eigenmap (LE) (Belkin and Niyogi, 2003). These methods are more suitable for preserving the local geometrical structure of the original data rather than obtaining discriminant features. That is because none of these algorithms can obtain a low-dimensional representation of a new test sample, which is the so-called ‘out of sample’ problem. One common way to deal with this problem is to construct an explicit linear mapping between the high-dimensional data and the low-dimensional subspace. Locality preserving projection (LPP) (He and Niyogi, 2003) is a typical representation method of this approach. However, LPP does not take the label information of samples into account. Thus, more recent studies have made use of label information and derived a lot of supervised feature extraction algorithms, such as supervised optimal locality preserving projection (SOLPP) (Wong and Zhao, 2012) and local linear discriminant analysis (LLDA) (Fan *et al.*, 2011).

Yan *et al.* (2007) presented a general formulation called graph embedding, and the above-mentioned algorithms, such as PCA, LDA, LPP, LLE, and the recently proposed tensor based algorithms, can all be reformulated within this common framework. In graph embedding, two graphs should be constructed, i.e., the specific intrinsic graph $G = \{\mathbf{X}, \mathbf{W}\}$ that describes certain enhanced statistical or geometric properties of the original data, and the penalty graph $G_P = \{\mathbf{X}, \mathbf{W}_P\}$ that demonstrates statistical or geometric properties of the original data that should be constrained. Marginal Fisher analysis (MFA) was developed using the graph embedding framework (Yan *et al.*, 2007). Another popular algorithm based on graph embedding is local discriminant embedding (LDE) (Chen *et al.*, 2005). Note that LDE and MFA are essentially the same.

The aforementioned algorithms also achieve good classification accuracy through experiments. However, an important property named ‘statistical uncorrelation’ is ignored or not maintained. The statistical uncorrelated criterion aims to eliminate the redundancy among the extracted discriminant fea-

tures. Jin *et al.* (2001) introduced an uncorrelated optimal discriminant vector (UODV) and a related theorem. To cope with the disadvantage of UODV, which cannot reflect the total scatter of the whole sample set, Jing *et al.* (2003) proposed an improved uncorrelated optimal discriminant vector (IUODV). However, they considered only a globally statistical uncorrelated criterion over all data. Thus, a local uncorrelated discriminant transform (LUODT) was further recommended, and a local uncorrelated criterion via reconstructing the total scatter matrix by redefining the expectation of each data point was proposed (Jing *et al.*, 2011). However, LUODT cannot take the class information of data into account, which is crucial to classification. Chen *et al.* (2013) recommended the local uncorrelated discriminant projection (LUDP) depending on a reformative local uncorrelated criterion and imposed this new constraint into maximum margin analysis. Overall, all these methods fail to capture the local information of interclass and intraclass simultaneously.

To resolve this problem, a novel feature extraction method called local uncorrelated local discriminant embedding (LULDE) is proposed in this paper. There are several contributions in the proposed approach. First, we propose a new local uncorrelated criterion, which effectively preserves both the local information of interclass and the local information of intraclass. Second, we reconstruct the affinity matrices of graphs G and G_P , which makes our algorithm obtain more discriminant capacity than the conventional LDE algorithm. Finally, the proposed method uses a different approach from PCA to preprocess the high-dimensional data, which overcomes the small-sample-size (SSS) problem.

2 Review of local discriminant embedding

Suppose that there is a training matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathbb{R}^m$ ($i = 1, 2, \dots, n$) and m is the dimensionality of the sample point. Each of them belongs to one of c classes $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_c\}$. The number of training samples in class i , n_i , satisfies $\sum_{i=1}^c n_i = n$. The class label of data point \mathbf{x}_i is denoted by $l_{\mathbf{x}_i}$. In this section, we will give a brief review of LDE.

LDE (Chen *et al.*, 2005) aims to find a

low-dimensional embedding such that data points of the same class maintain their intrinsic neighbor relationships, whereas neighboring points of different classes keep away from each other.

To discover both geometrical and discriminant structures of the data manifold, two graphs are built: intrinsic graph G and penalty graph G_P . Assume that the sets of intra-class and inter-class neighbors of \mathbf{x}_i are indicated by $\text{NN}_I(\mathbf{x}_i)$ and $\text{NN}_E(\mathbf{x}_i)$, respectively. We have

$$\text{NN}_I(\mathbf{x}_i) = \{\mathbf{x}_j | l_{\mathbf{x}_i} = l_{\mathbf{x}_j} \text{ and } [\mathbf{x}_i \in N_{k_1}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_{k_1}(\mathbf{x}_i)]\}, \quad (1)$$

$$\text{NN}_E(\mathbf{x}_i) = \{\mathbf{x}_j | l_{\mathbf{x}_i} \neq l_{\mathbf{x}_j} \text{ and } [\mathbf{x}_i \in N_{k_2}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_{k_2}(\mathbf{x}_i)]\}, \quad (2)$$

where $N_k(\mathbf{x})$ denotes the k -nearest neighbors of \mathbf{x} . Let \mathbf{W} and \mathbf{W}_P denote the affinity matrices of graphs G and G_P , respectively. Each element of these matrices can be defined as follows:

$$W_{ij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}, & \mathbf{x}_i \in \text{NN}_I(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \text{NN}_I(\mathbf{x}_i), \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

$$W_{Pij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}, & \mathbf{x}_i \in \text{NN}_E(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \text{NN}_E(\mathbf{x}_i), \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where t is a constant, usually set to the average of squared distances between all pairs. Thus, the optimization problem of LDE is as follows:

$$\begin{aligned} J(\mathbf{V}) &= \arg \max \sum_{i,j} \|\mathbf{V}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{x}_j\|^2 W_{Pij} \\ \text{s.t. } & \sum_{i,j} \|\mathbf{V}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{x}_j\|^2 W_{ij} = 1. \end{aligned} \quad (5)$$

Using simple matrix algebra, the aforementioned objective function can be reformulated as

$$\begin{aligned} J(\mathbf{V}) &= \arg \max \frac{\text{tr}\{\mathbf{V}^T \mathbf{X}(\mathbf{D}_P - \mathbf{W}_P)\mathbf{X}^T \mathbf{V}\}}{\text{tr}\{\mathbf{V}^T \mathbf{X}(\mathbf{D} - \mathbf{W})\mathbf{X}^T \mathbf{V}\}} \\ &= \arg \max \frac{\text{tr}(\mathbf{V}^T \mathbf{X} \mathbf{L}_P \mathbf{X}^T \mathbf{V})}{\text{tr}(\mathbf{V}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{V})}, \end{aligned} \quad (6)$$

where $\text{tr}(\mathbf{W})$ denotes the trace of matrix \mathbf{W} , \mathbf{D} and \mathbf{D}_P are diagonal matrices, whose entries are column (or row, since \mathbf{W} and \mathbf{W}_P are symmetric matrices)

sums of \mathbf{W} and \mathbf{W}_P , respectively, and \mathbf{L} and \mathbf{L}_P denote the Laplacian matrices associated with graphs G and G_P , respectively.

Then we obtain a projection matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d]$ whose columns are the generalized eigenvectors corresponding to the d largest eigenvalues of the equation

$$\mathbf{X} \mathbf{L}_P \mathbf{X}^T \mathbf{v}_j = \lambda \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{v}_j. \quad (7)$$

3 Local uncorrelated local discriminant embedding

3.1 Statistical uncorrelation

The statistical uncorrelation is a very important property, and the extracted factors are ignored or not preserved in some algorithms. This property is aimed to eliminate the redundancy among the extracted discriminant features and ensure that the extracted discriminant projection vectors have no overlapping discriminant information from the statistical point of view.

The statistical uncorrelation is defined as follows (Jin *et al.*, 2001):

$$\mathbf{v}_j^T \mathbf{S}_T \mathbf{v}_i = 0 \quad (i = 1, 2, \dots, j-1), \quad (8)$$

where \mathbf{S}_T denotes the total scatter matrix satisfying

$$\mathbf{S}_T = \sum_{i,j=1}^n (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \quad (9)$$

From the aforementioned equation, we can find that the classical statistical uncorrelation constraint is a globally statistical uncorrelation. However, the local information is more important than the global. Recently, some local uncorrelated approaches were introduced by extracting the local information of the data (Jing *et al.*, 2011; Chen *et al.*, 2013). However, these algorithms cannot use the local information of intra-class and inter-class simultaneously.

In this study, we try to reconstruct the total scatter matrix by redefining a local matrix $\mathbf{L}\mathbf{L}$, and construct the reformative uncorrelated constraints.

3.2 Local statistical uncorrelation

First, we construct a matrix $\mathbf{L}\mathbf{L}$ that preserves the local information about the same class and the

different classes. The entry of $\mathbf{L}\mathbf{L}$ is defined as follows:

$$LL_{ij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}(1 + e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}), & \mathbf{x}_i \in \text{NN}_I(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \text{NN}_I(\mathbf{x}_i), \\ e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}(1 - e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}), & \mathbf{x}_i \in \text{NN}_E(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \text{NN}_E(\mathbf{x}_i), \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

where $\text{NN}_I(\mathbf{x}_i)$ and $\text{NN}_E(\mathbf{x}_i)$ have been defined in Eqs. (1) and (2), respectively. Then we redefine the total scatter matrix \mathbf{S}_L as follows:

$$\mathbf{S}_L = \sum_{i,j=1}^n LL_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \quad (11)$$

Hence, we define a new local uncorrelated constraint as follows:

$$\mathbf{v}_j^T \mathbf{S}_L \mathbf{v}_i = 0 \quad (i = 1, 2, \dots, j - 1). \quad (12)$$

3.3 Local uncorrelated local discriminant embedding

To use the local information about the same class and the different classes, we will introduce the proposed LULDE, which is derived from the graph embedding framework. First, we redefine the affinity matrices $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}_P$ as follows:

$$\widetilde{W}_{ij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}(1 + e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}), & \mathbf{x}_i \in \text{NN}_I(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \text{NN}_I(\mathbf{x}_i), \\ 0, & \text{otherwise,} \end{cases} \quad (13)$$

$$\widetilde{W}_{Pij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}(1 - e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}), & \mathbf{x}_i \in \text{NN}_E(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \text{NN}_E(\mathbf{x}_i), \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

Thus, the optimization problem of LULDE is as follows:

$$J(\mathbf{V}) = \arg \max \frac{\sum_{i,j} \|\mathbf{V}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{x}_j\|^2 \widetilde{W}_{Pij}}{\sum_{i,j} \|\mathbf{V}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{x}_j\|^2 \widetilde{W}_{ij}} \\ \text{s.t. } \mathbf{V}^T \mathbf{V} = \mathbf{I}, \mathbf{v}_j^T \mathbf{S}_L \mathbf{v}_i = 0, i = 1, 2, \dots, j - 1, \quad (15)$$

where \mathbf{I} is the identity matrix. Using simple matrix algebra, the aforementioned objective function can

be reformulated as

$$J(\mathbf{V}) = \arg \max \frac{\text{tr}\{\mathbf{V}^T \mathbf{X}(\widetilde{\mathbf{D}}_P - \widetilde{\mathbf{W}}_P)\mathbf{X}^T \mathbf{V}\}}{\text{tr}\{\mathbf{V}^T \mathbf{X}(\widetilde{\mathbf{D}} - \widetilde{\mathbf{W}})\mathbf{X}^T \mathbf{V}\}} \\ = \arg \max \frac{\text{tr}(\mathbf{V}^T \mathbf{X} \widetilde{\mathbf{L}}_P \mathbf{X}^T \mathbf{V})}{\text{tr}(\mathbf{V}^T \mathbf{X} \widetilde{\mathbf{L}} \mathbf{X}^T \mathbf{V})} \\ = \arg \max \frac{\text{tr}(\mathbf{V}^T \mathbf{S}_P \mathbf{V})}{\text{tr}(\mathbf{V}^T \mathbf{S} \mathbf{V})}, \quad (16)$$

where $\mathbf{S}_P = \mathbf{X} \widetilde{\mathbf{L}}_P \mathbf{X}^T$, $\mathbf{S} = \mathbf{X} \widetilde{\mathbf{L}} \mathbf{X}^T$, $\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{D}}_P$ are the redefined diagonal matrices, whose entries are column (or row, since $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}_P$ are symmetric matrices) sums of the redefined $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}_P$, respectively, and $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{L}}_P$ denote the Laplacian matrices associated with graphs G and G_P , respectively.

The first discriminant vector, \mathbf{v}_1 , which is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{S}^{-1} \mathbf{S}_P$, can be easily obtained. Then according to the following theorem, we can calculate the other $d - 1$ optimal discriminant vectors iteratively:

Theorem 1 If \mathbf{S} and \mathbf{S}_L are nonsingular matrices, then \mathbf{v}_j ($j \geq 2$) is the eigenvector corresponding to the maximum eigenvalue of the following equation:

$$\mathbf{P} \mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j = \lambda \mathbf{v}_j, \quad (17)$$

where

$$\begin{cases} \mathbf{P} = \mathbf{I} - \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d (\mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d)^{-1} \mathbf{V}_d^T \mathbf{S}_L, \\ \mathbf{V}_d = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{j-1}]. \end{cases} \quad (18)$$

Proof First, we use the Lagrange multiplier method to transform Eq. (16) including the local uncorrelated constraint

$$L(\mathbf{v}_j) = \mathbf{v}_j^T \mathbf{S}_P \mathbf{v}_j - \lambda \mathbf{v}_j^T \mathbf{S} \mathbf{v}_j - \sum_{i=1}^{j-1} u_i \mathbf{v}_j^T \mathbf{S}_L \mathbf{v}_i. \quad (19)$$

Letting the partial derivatives $\partial L(\mathbf{v}_j) / \partial \mathbf{v}_j$ equal zero and $\mathbf{u} = [u_1, u_2, \dots, u_{j-1}]$, we have

$$\frac{\partial L(\mathbf{v}_j)}{\partial \mathbf{v}_j} = 2\mathbf{S}_P \mathbf{v}_j - 2\lambda \mathbf{S} \mathbf{v}_j - \mathbf{S}_L \mathbf{V}_d \mathbf{u} = 0. \quad (20)$$

Left multiplying Eq. (20) by $\mathbf{v}_i^T \mathbf{S}_L \mathbf{S}^{-1}$ ($i = 1, 2, \dots, j - 1$), we obtain a set of $j - 1$ equations as follows:

$$2\mathbf{v}_i^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j - 2\lambda \mathbf{v}_i^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S} \mathbf{v}_j \\ - \mathbf{v}_i^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d \mathbf{u} = 0 \quad (i = 1, 2, \dots, j - 1). \quad (21)$$

Eq. (21) can be represented in the form of matrix:

$$2\mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j - \mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d \mathbf{u} = 0. \quad (22)$$

Therefore, we obtain

$$\mathbf{u} = 2(\mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d)^{-1} \mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j. \quad (23)$$

Left multiplying Eq. (20) by \mathbf{S}^{-1} , we can obtain

$$2\mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j - 2\lambda \mathbf{v}_j - \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d \mathbf{u} = 0. \quad (24)$$

Substituting Eq. (23) into Eq. (24) leads to

$$2\mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j - 2\lambda \mathbf{v}_j - 2\mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d \cdot (\mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d)^{-1} \mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j = 0, \quad (25)$$

i.e.,

$$(\mathbf{I} - \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d (\mathbf{V}_d^T \mathbf{S}_L \mathbf{S}^{-1} \mathbf{S}_L \mathbf{V}_d)^{-1} \mathbf{V}_d^T \mathbf{S}_L) \cdot \mathbf{S}^{-1} \mathbf{S}_P \mathbf{v}_j = \lambda \mathbf{v}_j.$$

3.4 Small-sample-size (SSS) problem

In many real face recognition problems, the dimension of data is far greater than the number of samples, which causes the matrix based on the data to be singular. This is the well-known SSS problem. When suffering from the SSS problem, many algorithms cannot continue to work. To deal with this problem, many methods have been proposed usually applying PCA to reduce the dimensionality of the original data. However, the PCA method simply abandons the null space of the covariance matrix of the training set, and this may lose some discriminant information. In this study, the proposed algorithm can avoid PCA.

We can see that the redefined matrix \mathbf{S}_L is the sum of matrices \mathbf{S} and \mathbf{S}_P , i.e., $\mathbf{S}_L = \mathbf{S} + \mathbf{S}_P$. Hence, we remove the null space of \mathbf{S}_L , and project the samples into its range space. Suppose $\mathbf{U} \in \mathbb{R}^{m \times d}$ ($d \ll m$) is the projection matrix. Then these matrices can be expressed as $\mathbf{S}_L = \mathbf{U}^T \mathbf{S}_L \mathbf{U}$, $\mathbf{S} = \mathbf{U}^T \mathbf{S} \mathbf{U}$, and $\mathbf{S}_P = \mathbf{U}^T \mathbf{S}_P \mathbf{U}$. Thus, we make these matrices nonsingular and the computation is then clearly reduced and the SSS problem can be overcome. The whole procedure of LULDE is summarized in Algorithm 1.

4 Experiments and analysis

In this section, we evaluate the effectiveness of the proposed LULDE method. The performance of

Algorithm 1 Local uncorrelated local discriminant embedding (LULDE)

Input: training set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ for c classes, and the desired final dimensionality d .

Output: optimal projection matrix \mathbf{V} .

- 1: Construct the affinity matrices $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{W}}_P$ defined in Eqs. (13) and (14).
 - 2: Construct the local uncorrelated matrix \mathbf{LL} defined in Eq. (10).
 - 3: Compute \mathbf{S}_L according to Eq. (11) and \mathbf{S} and \mathbf{S}_P according to Eq. (16).
 - 4: Compute the projection matrix \mathbf{U} which is the eigenvector corresponding to the nonzero eigenvalue of \mathbf{S}_L .
 - 5: Project \mathbf{S}_L , \mathbf{S} , and \mathbf{S}_P to the range space of \mathbf{S}_L , $\mathbf{S}_L = \mathbf{U}^T \mathbf{S}_L \mathbf{U}$, $\mathbf{S} = \mathbf{U}^T \mathbf{S} \mathbf{U}$, and $\mathbf{S}_P = \mathbf{U}^T \mathbf{S}_P \mathbf{U}$.
 - 6: Compute the first projection vector, \mathbf{v}_1 , which is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{S}^{-1} \mathbf{S}_P$.
 - 7: Compute the other projection vector \mathbf{v}_j ($j \geq 2$) which is the eigenvector corresponding to the maximum eigenvalue of Eq. (17), until $j > d$.
 - 8: The d projection vectors construct the projection matrix $\mathbf{V}_{LULDE} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d]$.
 - 9: Obtain the optimal projection matrix $\mathbf{V} = \mathbf{U} \mathbf{V}_{LULDE}$.
-

LULDE was compared to those of LDA, IUODV, LUDT, LUDP, and LDE on four databases, namely, Yale, ORL, Extended Yale B, and FERET. To overcome the SSS problem, we applied PCA to preprocess the data before implementing the feature extraction algorithm (IUODV, LUDT, LDE). About 99% of variance was retained in our experiments. For LUDT and LUDP, we set the number of the nearest neighbors (k_1) between 1 and 10 and chose the best accuracy as the final result. For LDE and LULDE algorithms, we fixed $k_1 = 5$ and $k_2 = 10$ (also set $k_2 = 20$ in big databases, e.g., Extended Yale B and FERET, as a contrast), and set the heat parameter t to the average of squared distances between all pairs about databases. In the experiments, the nearest neighbor (NN) classifier with the Euclidean distance metric was employed as the classification algorithm. All experiments were repeated 10 times with different training samples and the average accuracy rates were recorded.

4.1 Experiments on the Yale database

In the Yale face database, there are 165 gray images of 15 persons, and each individual has 11 images with a resolution of 92×112 pixels. The images of the cropped version contain lighting variations and facial expression variations (normal, happy, sad, sleepy, surprised, and winking). In our experiment, each image was manually cropped and resized to 32×32 pixels. Some images from one person are shown in Fig. 1.



Fig. 1 Images of one person from the Yale database

In this experiment, we randomly chose r ($r = 2, 3, \dots, 8$) images for each person as training samples, and the remainder for testing. Table 1 lists the best average recognition rates and the corresponding standard deviations of the different methods with different training samples. It is observed that the recognition rates of all methods improved significantly as the number of training samples increased. That is because more information can be obtained with a large set of training data than with a small set. We also observe that the classification results of our proposed method were better than those of the other compared algorithms on all training subsets. Fig. 2 shows the recognition rate versus dimensionality, where we chose five training samples per person. Fig. 3 shows the recognition rate versus the number of training samples per class on the Yale database. Our proposed LULDE algorithm also consistently outperformed the other methods.

Finally, to investigate the performance of LUDT, LUDP, LDE, and LULDE methods under k_1 , we randomly selected eight images for each person as training samples, with the remaining images for testing. We fixed $k_2 = 10$ and set k_1 between 1 and 10. The experiments were repeated 10 times and the best average recognition rates were recorded in Table 2. Fig. 4 displays the recognition rate versus k_1 .

4.2 Experiments on the ORL database

The ORL face database includes 400 face images of 40 individuals, and has different variations including expression, lighting, and facial details. There are

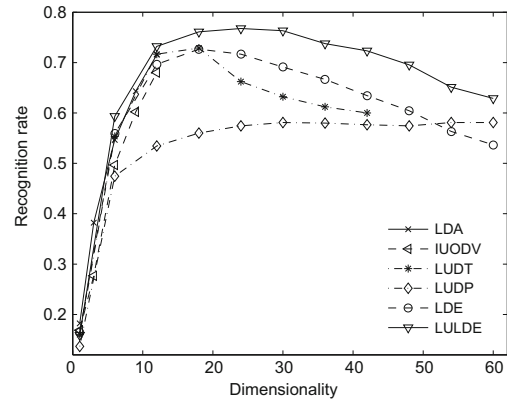


Fig. 2 Recognition rate versus dimensionality on the Yale database

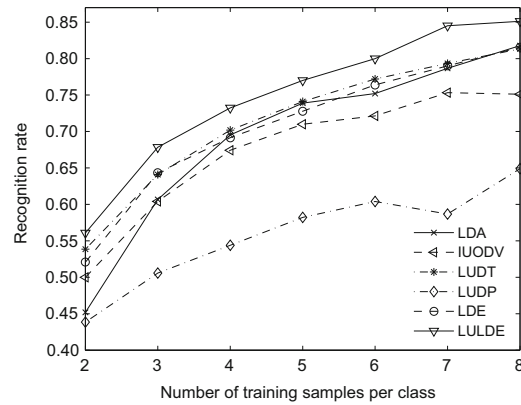


Fig. 3 Recognition rate versus the number of training samples on the Yale database

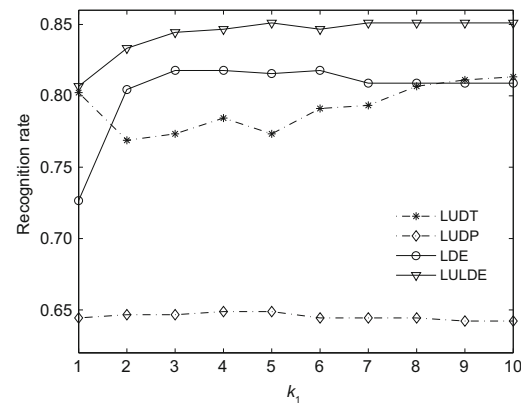


Fig. 4 Recognition rate versus k_1 on the Yale database

10 images for each subject, and the resolution of each image is 92×112 pixels. The images were resized to 32×32 pixels in our experiment. Fig. 5 shows several images of one person. We randomly grouped the images of each person into two parts. One part was

Table 1 Performance comparisons on the Yale database for different numbers of training samples per class

Algorithm	Maximum average recognition rate±standard deviation (%)						
	$r = 2$	3	4	5	6	7	8
LDA	45.19±3.14	60.67±5.57	69.52±4.44	73.89±5.52	75.20±2.20	78.67±3.58	81.78±5.52
IUODV	50.00±3.48	60.42±3.36	67.43±4.42	71.00±5.65	72.13±2.22	75.33±5.82	75.11±5.42
LUDT	53.85±4.80	64.08±3.92	70.19±3.01	74.11±4.80	77.20±3.17	79.33±4.98	81.33±5.66
LUDP	43.85±3.29	50.58±3.69	54.38±3.69	58.22±2.47	60.40±3.27	58.67±8.20	64.89±9.06
LDE	52.07±2.79	64.33±3.98	69.14±2.95	72.78±5.11	76.40±4.49	79.00±5.99	81.56±6.55
LULDE	56.07±2.20	67.83±3.47	73.24±3.25	77.00±2.92	80.00±3.82	84.50±4.65	85.11±5.83

Table 2 Results under different k_1 's on the Yale database

Algorithm	Best average recognition rate (%)									
	$k_1 = 1$	2	3	4	5	6	7	8	9	10
LUDT	80.22	76.89	77.33	78.44	77.33	79.11	79.33	80.67	81.11	81.33
LUDP	64.44	64.67	64.67	64.89	64.89	64.44	64.44	64.44	64.22	64.22
LDE	72.67	80.44	81.78	81.78	81.56	81.78	80.89	80.89	80.89	80.89
LULDE	80.67	83.33	84.44	84.67	85.11	84.67	85.11	85.11	85.11	85.11



Fig. 5 Images of one person from the ORL database

used as training samples with r ($r = 2, 3, \dots, 8$) images being chosen for each individual, and the other as testing samples.

Table 3 reports the maximum average recognition rates and the corresponding standard deviations of the different methods with different numbers of training samples. The proposed method (LULDE) achieved the best recognition rate compared with LDA, IUODV, LUDT, LUDP, and LDE. Fig. 6 illustrates the recognition rate versus the dimensionality, where we chose five training samples per person. The recognition rate versus different training samples per class is shown in Fig. 7. Figs. 6 and 7 show that the proposed method (LULDE) also consistently outperformed the other methods in most experimental cases.

To investigate the capability of the LUDT, LUDP, LDE, and LULDE methods with different numbers of nearest neighbors (k_1), we randomly selected eight images for each person as training samples, with the remaining images for testing. We fixed $k_2 = 10$ and set k_1 between 1 and 10. The experiments were repeated 10 times and the maximum average recognition rates were recorded in Table 4. Fig. 8 displays the recognition rate versus k_1 .

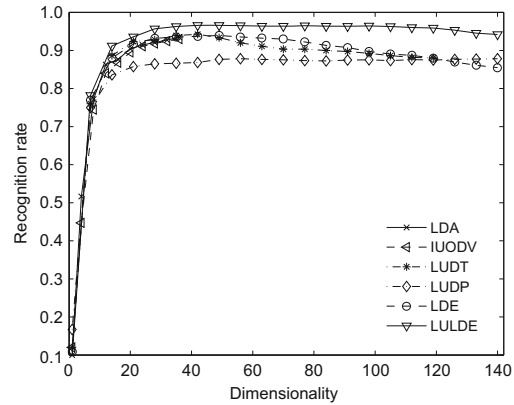


Fig. 6 Recognition rate versus dimensionality on the ORL database

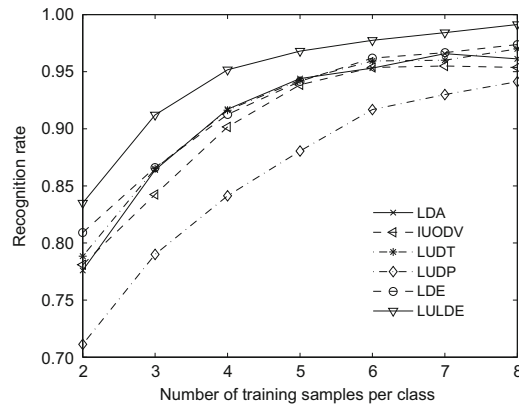


Fig. 7 Recognition rate versus the number of training samples per class on the ORL database

Table 3 Performance comparisons on the ORL database for different numbers of training samples per class

Algorithm	Maximum average recognition rate±standard deviation (%)						
	$r = 2$	3	4	5	6	7	8
LDA	77.59±3.32	86.39±2.64	91.71±1.84	94.40±2.48	95.31±1.96	96.58±1.94	96.13±1.24
IUODV	78.09±3.41	84.25±2.42	90.17±1.88	93.85±2.08	95.38±1.54	95.50±1.72	95.38±1.87
LU DT	78.84±2.92	86.50±1.59	91.63±2.19	94.25±2.44	95.94±1.51	96.00±1.41	97.00±1.05
LUDP	71.13±2.75	79.00±1.80	84.13±3.01	88.05±2.67	91.69±2.23	93.00±2.12	94.13±2.64
LDE	80.91±2.85	86.61±2.20	91.25±2.10	94.15±2.10	96.19±1.92	96.67±1.30	97.38±0.92
LULDE	83.50±2.68	91.21±1.76	95.17±0.97	96.80±1.83	97.75±1.56	98.42±1.33	99.13±0.60

Table 4 Results under different k_1 's on the ORL database

Algorithm	Best average recognition rate (%)									
	$k_1 = 1$	2	3	4	5	6	7	8	9	10
LU DT	97.00	95.25	95.13	95.50	95.75	95.87	96.13	96.38	96.63	96.38
LUDP	94.13	94.13	94.13	94.13	94.13	94.13	94.13	94.13	94.13	94.13
LDE	94.50	97.50	97.88	97.62	97.38	97.62	97.50	97.50	97.50	97.50
LULDE	98.75	99.13	99.13	98.88	99.13	99.13	99.13	99.13	99.13	99.13

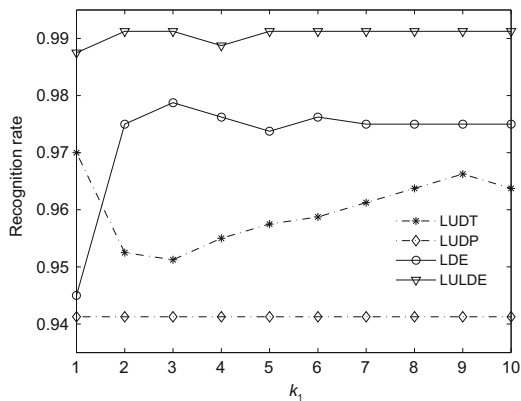


Fig. 8 Recognition rate versus k_1 on the ORL database

4.3 Experiments on the Extended Yale B database

The Extended Yale B face database contains many gray face images of 38 subjects under different pose and illumination conditions. In our experiment, we chose a subset of the database that includes only those under illumination conditions. Each person had 64 different front images. All the images were resized to a resolution of 32×32 pixels. Some images from one person are illustrated in Fig. 9.

We randomly chose r ($r = 5, 10$) images for each person as training samples, and the remainder for testing. Table 5 shows the maximum average recognition rates and the corresponding standard devia-

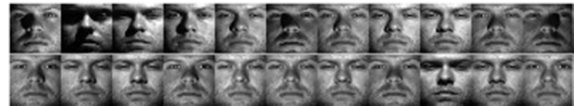


Fig. 9 Images of one person from the Extended Yale B database

Table 5 Performance comparisons on the Extended Yale B database for different numbers of training samples per class

Algorithm	Maximum average recognition rate ±standard deviation (%)	
	$r = 5$	$r = 10$
LDA	65.85±2.10	79.35±1.59
IUODV	65.23±2.12	79.17±1.50
LU DT	66.46±1.79	80.40±1.25
LUDP	30.77±0.93	43.51±1.15
LDE	65.62±1.87	79.88±1.42
LULDE	68.73±1.79	81.95±1.45

tions. The proposed method (LULDE) obtained the best recognition rate when the number of training samples for each class varied from 5 to 10. Fig. 10 illustrates the recognition rate versus the dimensionality, and the number of training samples per subject was 5. Fig. 10 shows that the LULDE algorithm achieved a higher recognition rate than LDA, IUODV, LU DT, LDUP, and LDE.

Also, to investigate the capability of the LU DT, LUDP, LDE, and LULDE methods under different numbers of nearest neighbors (k_1), we randomly selected 10 images for each person as training samples,

with the remaining images for testing. We set parameter k_1 between 1 and 10, fixed $k_2 = 10$ and also fixed $k_2 = 20$ as a contrast. The experiments were repeated 10 times and the best average recognition rates were recorded in Table 6. Fig. 11 shows the recognition rate versus k_1 .

4.4 Experiments on the FERET database

The FERET face database contains 200 individuals and each person has 7 images which include different illumination conditions and facial expressions. All the original gray images were resized to 32×32 pixels. Fig. 12 illustrates several images of one person. We randomly grouped the images of each person into two parts. One part was used as training samples with r ($r = 3, 4$) images being chosen for each individual, and the other as testing samples.

The maximum average recognition rates and the corresponding standard deviations of the different methods are shown in Table 7. The proposed LULDE algorithm achieved the best recognition rate compared with other methods when the number of training samples per subject was 3. How-

ever, when using 4 training samples per subject it did not achieve the best performance. Fig. 13 shows the recognition rate versus the dimensionality, where the number of training samples per class was 3. The recognition rate of LULDE was not as good as that of IUODV, LUDT, or LDE when the dimension

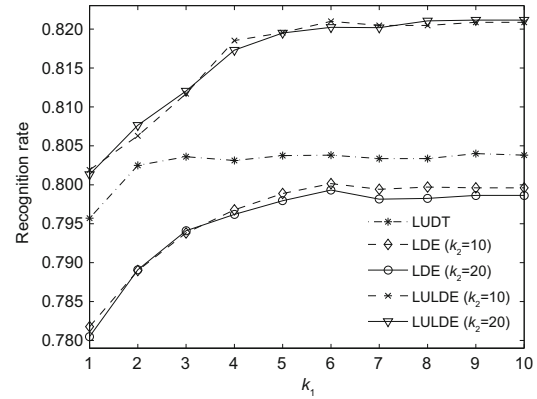


Fig. 11 Recognition rate versus k_1 on the Extended Yale B database



Fig. 12 Images of one person from the FERET database

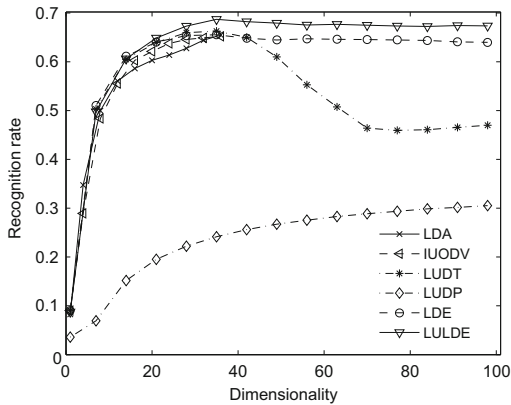


Fig. 10 Recognition rate versus dimensionality on the Extended Yale B database

Table 7 Performance comparisons on the FERET database for different numbers of training samples per class

Algorithm	Maximum average recognition rate \pm standard deviation (%)	
	$r = 3$	$r = 4$
LDA	37.56 \pm 1.95	35.48 \pm 1.90
IUODV	77.81 \pm 0.63	86.55 \pm 1.12
LUDT	76.76 \pm 1.44	83.97 \pm 1.82
LUDP	36.06 \pm 1.08	41.75 \pm 1.33
LDE	77.10 \pm 1.58	86.47 \pm 1.09
LULDE	77.88 \pm 1.32	85.23 \pm 0.97

Table 6 Results under different k_1 's on the Extended Yale B database

Algorithm	Best average recognition rate (%)									
	$k_1 = 1$	2	3	4	5	6	7	8	9	10
LUDT	79.57	80.25	80.36	80.31	80.38	80.38	80.34	80.34	80.40	80.38
LUDP	43.50	43.49	43.48	43.49	43.50	43.49	43.50	43.51	43.51	43.50
LDE ($k_2 = 10$)	78.17	78.90	79.38	79.68	79.89	80.02	79.94	79.97	79.96	79.96
LDE ($k_2 = 20$)	78.05	78.91	79.41	79.62	79.80	79.93	79.82	79.82	79.83	79.86
LULDE ($k_2 = 10$)	80.19	80.63	81.17	81.85	81.95	82.10	82.05	82.05	82.09	82.09
LULDE ($k_2 = 20$)	80.13	80.77	81.20	81.73	81.95	82.02	82.02	82.11	82.12	82.12

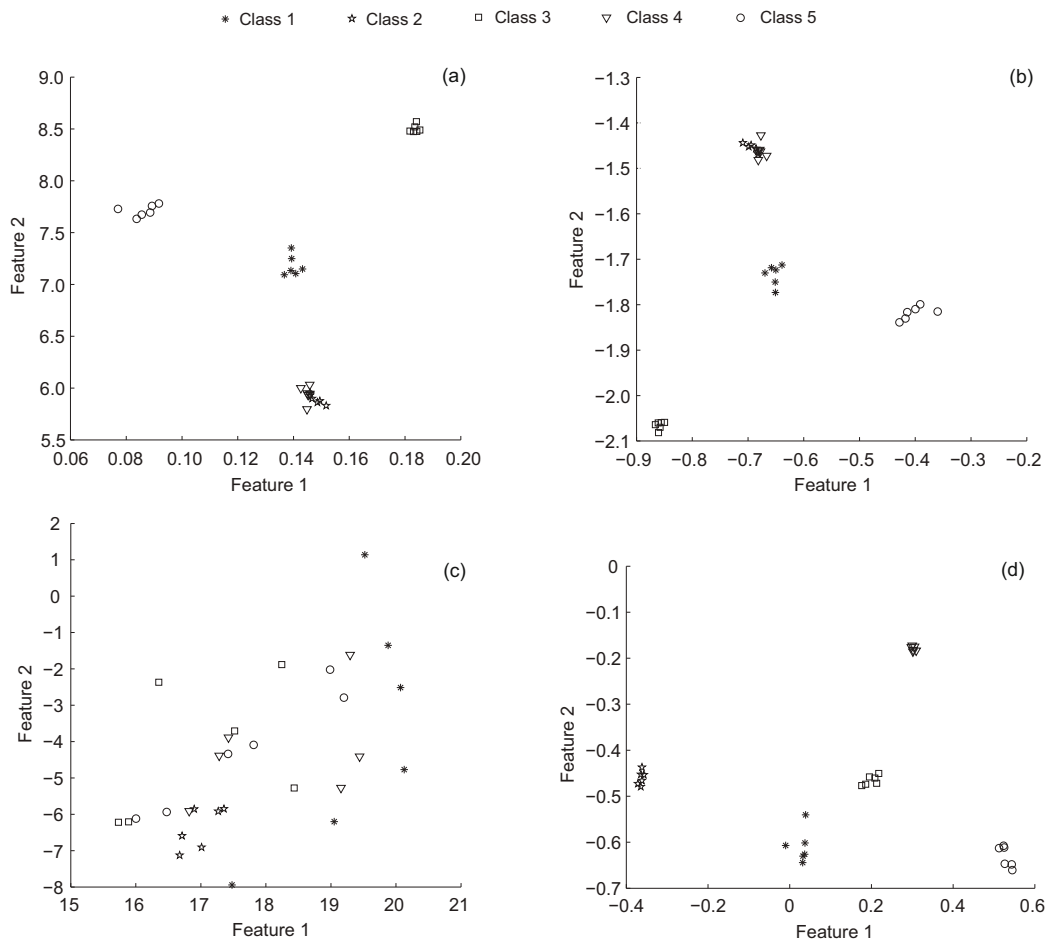


Fig. 15 ORL face images in a 2D Euclidean space using different algorithms: (a) IUODV; (b) LUDT; (c) LUDP; (d) LULDE

property. Finally, it overcomes the SSS problem without using PCA to preprocess the original data. Extensive experimental results on four face databases, Yale, ORL, Extended Yale B, and FERET, demonstrate the effectiveness of the proposed algorithm. However, choosing the number of nearest neighbors is also an open problem. Future work will be devoted to parameter selection and developing a nonparametric algorithm.

References

- Belhumeur, P.N., Hespanha, J.P., Kriegman, D., 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Patt. Anal. Mach. Intell.*, **19**(7):711-720. <http://dx.doi.org/10.1109/34.598228>
- Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neur. Comput.*, **15**(6):1373-1396.
- Chen, H.T., Chang, H.W., Liu, T.L., 2005. Local discriminant embedding and its variants. *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, p.846-853. <http://dx.doi.org/10.1109/CVPR.2005.216>
- Chen, Y., Zheng, W.S., Xu, X.H., et al., 2013. Discriminant subspace learning constrained by locally statistical uncorrelation for face recognition. *Neur. Netw.*, **42**:28-43. <http://dx.doi.org/10.1016/j.neunet.2013.01.009>
- Fan, Z.Z., Xu, Y., Zhang, D., 2011. Local linear discriminant analysis framework using sample neighbors. *IEEE Trans. Neur. Netw.*, **22**(7):1119-1132. <http://dx.doi.org/10.1109/TNN.2011.2152852>
- He, X.F., Niyogi, P., 2003. Locality preserving projections. *Proc. Advances in Neural Information Processing Systems*, p.327-334.
- Jin, Z., Yang, J.Y., Hu, Z.S., et al., 2001. Face recognition based on the uncorrelated discriminant transformation. *Patt. Recog.*, **34**(7):1405-1416. [http://dx.doi.org/10.1016/S0031-3203\(00\)00084-4](http://dx.doi.org/10.1016/S0031-3203(00)00084-4)
- Jing, X.Y., Zhang, D., Jin, Z., 2003. UODV: improved algorithm and generalized theory. *Patt. Recog.*,

- 36**(11):2593-2602.
[http://dx.doi.org/10.1016/S0031-3203\(03\)00177-8](http://dx.doi.org/10.1016/S0031-3203(03)00177-8)
- Jing, X.Y., Li, S., Zhang, D., *et al.*, 2011. Face recognition based on local uncorrelated and weighted global uncorrelated discriminant transforms. Proc. 18th IEEE Int. Conf. on Image Processing, p.3049-3052.
<http://dx.doi.org/10.1109/ICIP.2011.6116307>
- Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500):2323-2326.
<http://dx.doi.org/10.1126/science.290.5500.2323>
- Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500):2319-2323.
<http://dx.doi.org/10.1126/science.290.5500.2319>
- Turk, M., Pentland, A., 1991. Eigenfaces for recognition. *J. Cogn. Neurosci.*, **3**(1):71-86.
- Wong, W.K., Zhao, H.T., 2012. Supervised optimal locality preserving projection. *Patt. Recog.*, **45**(1):186-197.
<http://dx.doi.org/10.1016/j.patcog.2011.05.014>
- Yan, S.C., Xu, D., Zhang, B.Y., *et al.*, 2007. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Patt. Anal. Mach. Intell.*, **29**(1):40-51.
<http://dx.doi.org/10.1109/TPAMI.2007.250598>