

Discriminant Analysis with Local Gaussian Similarity Preserving for Feature Extraction

Xi Liu¹  · Zhengming Ma¹

Published online: 5 May 2017
© Springer Science+Business Media New York 2017

Abstract In this paper, we propose a novel discriminant analysis with local Gaussian similarity preserving (DA-LGSP) method for feature extraction. DA-LGSP can be viewed as a linear approximation of manifold learning based approach which seeks to find a linear projection that maximizes the between-class dissimilarities under the constraint of locality preserving. The local geometry of each point is preserved by the Gaussian coefficients of its neighbors, meanwhile the between-class dissimilarities are represented by Euclidean distances. Experiments are conducted on USPA data, COIL-20 dataset, ORL dataset and FERET dataset. The performance of the proposed method demonstrates that DA-LGSP is effective in feature extraction.

Keywords Feature extraction · Manifold learning · Fisher criterion · Between-class dissimilarities

1 Introduction

In recent years, feature extraction approaches for dimensionality reduction have received significant attention which is an important process in machine vision, pattern recognition task and exploratory data analysis [1]. Although feature extraction results in some loss of information about the original data, it retains meaningful features which have been demonstrated to be quite successful in biometrics [2], image retrieval [3,4], classification algorithms [5], and other areas [6].

Principal component analysis (PCA) is one of the most widely used data analysis tool for dimensionality reduction. Kernel PCA [7] is the extension of PCA to nonlinear dimensionality reduction and feature extraction. Nonnegative matrix factorization (NMF) and Manhattan NMF (MahNMF) [8] also are representative approach to dimensionality reduction having

✉ Xi Liu
liux239@mail2.sysu.edu.cn

¹ School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China

the property of learning parts-based representations [9,10]. Liu et al. propose large-cone NMF (LCNMF) [11] algorithms to obtain an attractive local solution for NMF. Another hot topic on dimensionality reduction is manifold learning for discovering the underlying meaningful low-dimensional structure hidden in the topology of high-dimensional nonlinear data set. Many algorithms have been developed including locally linear embedding (LLE) [12], ISOMAP [13], laplacian eigenmap (LE) [14], local tangent space alignment (LTSA) [15], Hessian LLE (HLLE) [16] and other extensions [17–27].

In order to agree to the task of pattern classification, projection techniques based on manifold learning are particularly suitable as a pretreatment step to classification. Locality preserving projection (LPP) [28] projects the original data into a subspace which preserves the local neighborhood structures by an optimal linear map. Analogous to LPP, unsupervised discriminant projection (UDP) [29] characterizes both the local scatter and the nonlocal scatter, seeking to find an optimal projection for globally maximizing and locally minimizing.

Most of the aforementioned algorithm are unsupervised methods, such as PCA, LLE, LPP and UDP. According to the quantity of supervised information used, existing dimensionality reduction methods can also be roughly categorized into unsupervised, semi-supervised and supervised methods. Semi-supervised learning (SSL) [30] exploits both unlabeled and labeled samples, high-order distance-based multiview stochastic learning (HD-MSL) is a semi-supervised image classification algorithm which improves hypergraph learning by simultaneously learning multiview features under a probabilistic framework. A branch of supervised learning is weakly supervised learning, in which the preference relationship between examples is indicated by weak cues [31]. Such as click feature could be regarded as weak cues for weakly supervised learning which have been successfully applies to image retrieval and image ranking [3,32]. This paper focuses on supervised learning is similar to Linear discriminant analysis (LDA). LDA may obtain good classification results since it takes full consideration of class labels. Ignoring class labels can result in misclassification of similar forms of different patterns, since the discrimination between a left pose image and a right pose image of one single person may be greater than the discrimination between two left pose images of two people. Face images of different persons should lie on corresponding manifolds but not a single manifold. This poses a problem that might be called “classification-oriented multi-manifolds learning” [28]. Multi-manifold learning assumes the data points lie on multiple underlying manifolds which are intersected or well separated. In order to achieve an optimal classification result, the low-dimensional embeddings corresponding to different manifolds should be as separable as possible in the final projected subspace.

LDA does not perform well when the data are non-Gaussian, since LDA is under the assumption of homoscedastic Gaussian class-conditional distributions. Constrained maximum variance mapping (CMVM) [33] takes local geometry and manifold labels into account. However, it ignores class labels when characterizing within-class scatter. Multi-manifold discriminant analysis (MMDA) [34] utilizes between-class Laplacian matrix and within-class Laplacian matrix to construct between-class graph and within-class graph. Kernel maximum distance analysis (KMMDA) maximizes the minimum distance of all class pairs in the low-dimensional subspace, and solves optimization problem by using the kernel trick [35]. Discriminative multi-manifold analysis (DMMA) seeks an optimal projection via inter-manifold graph and intra-manifold graph [36]. Both MMDA and DMMA serve as effective feature extraction algorithms for supervised classification tasks, but local topologic structures have not been fully considered. Nonparametric discriminant multi-manifold learning (NDML) [37] adopts LLE to preserve local geometry, and models separabilities between classes by manifold distance, while the manifold distance defined in NDML just takes adjacent classes into account but not all the classes.

To address the issues with the methods mentioned above, we propose a supervised method for feature extraction to classify data points sampled from multiple separated or intersecting nonlinear manifolds that are embedded in high-dimensional space, called discriminant analysis with local Gaussian similarity preserving (DA-LGSP). Our basic idea is to separate different classes farther under the constraint of local topological structures preserving. It is worthwhile to highlight several aspects of our method.

(1) We introduce a novel locality preserving method by Gaussian coefficients under the framework of thinking globally and fitting locally. The Gaussian coefficients can get good locality preserving effect.

(2) DA-LGSP explicitly considers class labels to preserve local geometry and construct between-class dissimilarities which are directly related to classification and recognition.

The rest of the paper is organized as follows: In Sect. 2, LDA and LLE are briefly reviewed. In Sect. 3, we describe the proposed algorithm in detail. In Sect. 4, the proposed algorithm is demonstrated on four datasets, and some discussions about the experimental results are also given. Section 5 finishes this paper with some conclusions.

2 Outline of LDA and LLE

2.1 LDA

The goal of linear discriminant analysis (LDA) is to project high-dimensional space to optimal discriminant vector space based linear projection such that Fisher criterion (i.e. the ratio of the between-class scatter to the within-class scatter) is maximized. In general, given a data set with training samples $X = [x_1 \dots x_N] \in R^{D \times N}$ and N is the total number of training samples, class labels are denoted by $z_i \in \{1, 2, \dots, c\}$ where c is the number of classes. We get their low-dimensional embedding $Y = [y_1 \dots y_N] \in R^{d \times N}$ by the projection axis W , where typically, $d < D$. The local between-class S_B and within-class S_W scatter matrices is defined as

$$S_B = \frac{1}{N} \sum_{i=1}^c N_i(m_i - m_o)(m_i - m_o)^T, \tag{1}$$

$$S_W = \frac{1}{N} \sum_{i=1}^c \sum_{x_k \in class(i)} (x_k - m_i)(x_k - m_i)^T, \tag{2}$$

$$S_T = \frac{1}{N} \sum_{i=1}^N (x_i - m_o)(x_i - m_o)^T. \tag{3}$$

Here $m_o = \frac{1}{N} \sum_{i=0}^N x_i$ is the mean vector of all training data, N_i is the number of training samples for the i th class, $\sum_{i=1}^c N_i = N$ and $m_i = \frac{1}{N_i} \sum_{x_i \in class(i)} x_i$ is the mean vector correspond to the i th class.

Both S_B and S_W are nonnegative definite matrix. The Fisher criterion is defined by

$$J(W) = \frac{W^T S_B W}{W^T S_W W}. \tag{4}$$

The optimal projection W is the generalized eigenvectors of $S_B W = \lambda S_W W$ corresponding to the d largest eigenvalues.

2.2 LLE

LLE is one of the typically manifold learning algorithms which the local geometrical information is explored and collected together to obtain a global optimum. It could well preserve local structure since it is thinking globally and fitting locally. The steps can be summarized as follows.

Step 1 For each data point x_i , identify its K -nearest-neighbors (KNN) in X with Euclidean distance metric, and note as $\tilde{X}_i = [x_i \ x_{i_1} \ \dots \ x_{i_K}] \in R^{D \times (K+1)}$.

Step2 Linearly reconstruct x_i with its KNN

$$\left\| x_i - \sum_{j=1}^K w_{ij} x_{i_j} \right\|_{w_i}^2 = \min, \tag{5}$$

where $w_i = [w_{i1} \ \dots \ w_{iK}]^T \in R^K$ and $\sum_{j=1}^K w_{ij} = 1$.

Minimizing the reconstruction error of x_i we can get

$$w_i = \frac{(\tilde{X}_i^T \tilde{X}_i)^{-1} \Gamma_K}{\Gamma_K^T (\tilde{X}_i^T \tilde{X}_i)^{-1} \Gamma_K}, \tag{6}$$

where $\tilde{X}_i = [x_{i_1} - x_i \ \dots \ x_{i_K} - x_i] \in R^{D \times K}$, $\Gamma_K = [1, 1, \dots, 1]^T \in R^K$.

Step 3 Linearly reconstruct the low-dimensional coordinates y_i with the same weights

$$\begin{aligned} \left\| y_i - \sum_{j=1}^K w_{ij} y_{i_j} \right\|^2 &= \left\| [y_i \ y_{i_1} \ \dots \ y_{i_K}] \begin{bmatrix} 1 \\ -w_{i1} \\ \vdots \\ -w_{iK} \end{bmatrix} \right\|^2 = \left\| Y_i \begin{bmatrix} 1 \\ -w_i \end{bmatrix} \right\|^2 \\ &= \text{tr} \left(Y_i \begin{bmatrix} 1 \\ -w_i \end{bmatrix} [1 \ -w_i^T] Y_i^T \right) = \text{tr} \left(Y_i \begin{bmatrix} 1 & -w_i^T \\ -w_i & w_i w_i^T \end{bmatrix} Y_i^T \right) \\ &= \text{tr} \left(Y_i A_i Y_i^T \right), \end{aligned} \tag{7}$$

where tr is the trace operator of matrix and $A_i = \begin{bmatrix} 1 & -w_i^T \\ -w_i & w_i w_i^T \end{bmatrix} \in R^{(K+1) \times (K+1)}$, note that

$$A_i = A_i^T.$$

Calculate the low-dimensional embedding Y for the N data points in X

$$\begin{aligned} \sum_{i=1}^N \left\| y_i - \sum_{j=1}^K w_{ij} y_{i_j} \right\|^2 &= \sum_{i=1}^N \text{tr} \left(Y_i A_i Y_i^T \right) = \text{tr} \left(\sum_{i=1}^N Y S_i A_i S_i^T Y^T \right) \\ &= \text{tr} \left(Y \left(\sum_{i=1}^N S_i A_i S_i^T \right) Y^T \right) = \text{tr} \left(Y A Y^T \right)_{Y Y^T = I_d} = \min, \end{aligned} \tag{8}$$

where $S_i \in \{0, 1\}^{N \times (K+1)}$ is a column selection matrix such that $Y_i = Y S_i$ and $A = \sum_{i=1}^N S_i A_i S_i^T$. Again, $A = A^T$.

Then the solution of Y is given by the eigenvector with the smallest nonzero eigenvalue.

3 DA-LGSP

Formally, given a data matrix with N data points $X = [x_1^{(1)}, x_2^{(1)}, \dots, x_i^{(p)}, \dots, x_N^{(c)}] \in R^{D \times N}$, which lie on different classes $\{M^{(1)}, M^{(2)}, \dots, M^{(c)}\}$, where N is the total number of training samples and c is the number of classes, for each point $x_i^{(p)}$ in X is a D -dimensional feature vector and p is the class which $x_i^{(p)}$ ($1 \leq p \leq c$) belongs to. DA-LGSP seeks to find a set of manifold coordinates $Y = [y_1^{(1)}, y_2^{(1)}, \dots, y_i^{(p)}, \dots, y_N^{(c)}] \in R^{d \times N}$ through a feature mapping $W: y_i^{(p)} = W^T x_i^{(p)}$, where typically, $d < D$. As discussed in Sect. 1, the optimal projection is found to preserve the local geometry and separate different classes apart.

3.1 Locality Structure

It is common believed that local feature space formed by nearest neighbors. Unsupervised local manifold learning approaches search the neighbors of a given point by applying KNN or ϵ -ball criterion, whereas DA-LGSP identify only the neighbors that are of the same class as the given point, which makes our methods more attractive for classification. Based on local linear fits, the local property of each neighborhood is represented by Guassian coefficients that best reconstruct each data point from the nearest neighborhood.

Defining X_i , the set of neighborhood nodes of node $x_i^{(p)}$ selected by supervise neighbor selection, $X_i = [x_i^{(p)} \ x_{i_1} \ \dots \ x_{i_K}] \in R^{D \times (K+1)}$, where $x_{i_j} \in M^{(p)}$. For simplicity, we neglect the class information p in this section since the neighbors are of the same class as $x_i^{(p)}$. Let the reconstruction weights of the neighbors of $x_i^{(p)}$ be $\rho_i = [\rho_{i1} \ \dots \ \rho_{iK}]^T \in R^K$, defined by

$$\begin{aligned}
 q_{ij} &= e^{-\frac{\|x_{i_j} - x_i\|^2}{2\sigma^2}}, \quad j = 1, \dots, K; \\
 \rho_{ij} &= \frac{q_{ij}}{\sum_{j=1}^K q_{ij}}, \quad j = 1, \dots, K;
 \end{aligned}
 \tag{9}$$

where q_{ij} is the Gaussian coefficient of x_{i_j} , and ρ_{ij} is the corresponding normalization coefficient, $\sum_{j=1}^K \rho_{ij} = 1$. In order to make ρ_{ij} more sensitive to distance, we choose the parameter σ as the average value of the distances between $x_i^{(p)}$ and its neighbors.

And then the low-dimensional coordinate $y_i^{(p)}$ of $x_i^{(p)}$ has been reconstructed with the Gaussian coefficients

$$y_i = \sum_{j=1}^K \rho_{ij} y_{i_j}.
 \tag{10}$$

Thus, we have a squared reconstruction error of $y_i^{(p)}$

$$\left\| y_i - \sum_{j=1}^K \rho_{ij} y_{i_j} \right\|^2 = \left\| [y_i \ y_{i_1} \ \dots \ y_{i_K}] \begin{bmatrix} 1 \\ -\rho_{i1} \\ \vdots \\ -\rho_{iK} \end{bmatrix} \right\|^2 = \left\| Y_i \begin{bmatrix} 1 \\ -\rho_i \end{bmatrix} \right\|^2$$

$$\begin{aligned}
 &= \text{tr} \left\{ Y_i \begin{bmatrix} 1 & -\rho_i^T \\ -\rho_i & \rho_i \rho_i^T \end{bmatrix} Y_i^T \right\} = \text{tr} \left\{ Y S_i \begin{bmatrix} 1 & -\rho_i^T \\ -\rho_i & \rho_i \rho_i^T \end{bmatrix} S_i^T Y^T \right\} \\
 &= \text{tr} \left\{ Y A_i Y^T \right\},
 \end{aligned} \tag{11}$$

where $S_i \in \{0, 1\}^{N \times (K+1)}$ is a column selection matrix such that $Y_i = Y S_i$, and $A_i = S_i \begin{bmatrix} 1 & -\rho_i^T \\ -\rho_i & \rho_i \rho_i^T \end{bmatrix} S_i^T$. Adding the squared construction error on N neighborhoods together

$$\begin{aligned}
 J_L &= \sum_{i=1}^N \left\| y_i - \sum_{j=1}^K \rho_{ij} y_{ij} \right\|^2 = \sum_{i=1}^N \text{tr} \left\{ Y A_i Y^T \right\} = \sum_{i=1}^N \text{tr} \left\{ W^T X A_i X^T W \right\} \\
 &= \text{tr} \left\{ W^T X \left(\sum_{i=1}^N A_i \right) X^T W \right\} = \text{tr} \left\{ W^T S_L W \right\},
 \end{aligned} \tag{12}$$

where $S_L = X \left(\sum_{i=1}^N S_i A_i S_i^T \right) X^T$.

3.2 Between-Class Dissimilarities

The Euclidean distance is often taken as a measure of dissimilarity. To some extent, large Euclidean distance between two points means high probability of their dissimilarities, otherwise they probably are similar to each other. So we define between-class dissimilarities derived from Euclidean distance to represent the dissimilarities of different classes. The between-class dissimilarities should be maximized in the projected subspace. We'll give the definition of between-class dissimilarities step by step.

To node $x_i^{(p)}$, $M^{(q)}$ is a class differ from $x_i^{(p)}$, i.e. $p \neq q$, the distance from point $x_i^{(p)}$ to class $M^{(q)}$ (denoted by $d(x_i^{(p)}, M^{(q)})$) is defined by

$$d(x_i^{(p)}, M^{(q)}) = \min_{x_j^{(q)} \in M^{(q)}} \left\| x_i^{(p)} - x_j^{(q)} \right\|^2 = \left\| x_i^{(p)} - n_{x_i}^{(q)} \right\|^2, \tag{13}$$

here $n_{x_i}^{(q)}$ is the nearest point to $x_i^{(p)}$ on $M^{(q)}$ which satisfies $n_{x_i}^{(q)} = \arg \min_{n_{x_i}^{(q)} \in M^{(q)}} \left(\left\| x_i^{(p)} - n_{x_i}^{(q)} \right\|^2 \right)$.

Next, it is important to define a measure of dissimilarity of two classes. The dissimilarity from class $M^{(p)}$ to class $M^{(q)}$ (denoted by $d(M^{(p)}, M^{(q)})$) is defined by

$$d(M^{(p)}, M^{(q)}) = \frac{1}{N_p} \sum_{x_i^{(p)} \in M^{(p)}} d(x_i^{(p)}, M^{(q)}), \tag{14}$$

where N_p is the number of training samples in class $M^{(p)}$. The dissimilarity between class $M^{(p)}$ and class $M^{(q)}$ (denoted by $D(M^{(p)}, M^{(q)})$) will be obtained as shown below

$$D(M^{(p)}, M^{(q)}) = d(M^{(p)}, M^{(q)}) + d(M^{(q)}, M^{(p)}). \tag{15}$$

Then the between-class dissimilarities in high-dimensional space can be defined as the following equation

$$\begin{aligned}
 J_D &= \sum_{p \neq q} D(M^{(p)}, M^{(q)}) = \sum_{p=1}^c \sum_{q=1, q \neq p}^c d(M^{(p)}, M^{(q)}) \\
 &= \sum_{p=1}^c \sum_{q=1, q \neq p}^c \frac{1}{N_p} \sum_{x_i^{(p)} \in M^{(p)}} d(x_i^{(p)}, M^{(q)}) \\
 &= \sum_{p=1}^c \sum_{x_i^{(p)} \in M^{(p)}} \sum_{q=1, q \neq p}^c \frac{1}{N_p} d(x_i^{(p)}, M^{(q)}) \\
 &= \sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \left\| x_i^{(p)} - n_{x_i}^{(q)} \right\|^2.
 \end{aligned} \tag{16}$$

Thus we obtain the between-class dissimilarities in the low-dimensional space through feature mapping W

$$\begin{aligned}
 J_D &= \sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \left\| y_i^{(p)} - n_{y_i}^{(q)} \right\|^2 \\
 &= \sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \text{tr} \left\{ \left(y_i^{(p)} - n_{y_i}^{(q)} \right) \left(y_i^{(p)} - n_{y_i}^{(q)} \right)^T \right\} \\
 &= \sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \text{tr} \left\{ \left(W^T x_i^{(p)} - W^T n_{x_i}^{(q)} \right) \left(W^T x_i^{(p)} - W^T n_{x_i}^{(q)} \right)^T \right\} \\
 &= \sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \text{tr} \left\{ W^T \left(x_i^{(p)} - n_{x_i}^{(q)} \right) \left(x_i^{(p)} - n_{x_i}^{(q)} \right)^T W \right\} \\
 &= \text{tr} \left\{ W^T \left(\sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \left(x_i^{(p)} - n_{x_i}^{(q)} \right) \left(x_i^{(p)} - n_{x_i}^{(q)} \right)^T \right) W \right\} \\
 &= \text{tr} \left\{ W^T S_D W \right\},
 \end{aligned} \tag{17}$$

where $S_D = \sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \left(x_i^{(p)} - n_{x_i}^{(q)} \right) \left(x_i^{(p)} - n_{x_i}^{(q)} \right)^T$

3.3 The Objective Function

In the proposed method, we expect to find the low-dimensional subspace obtained by an optimal projection W where different classes will be far located and locality will be well preserved.

$$\begin{aligned}
 J(W) &= \max(J_D) = \text{tr} \left\{ W^T S_D W \right\} \\
 \text{s.t. } & \text{tr} \left\{ W^T S_L W \right\} = \text{tr} \left\{ S_L \right\}
 \end{aligned} \tag{18}$$

$$\text{where } S_L = X \left(\sum_{i=1}^N S_i A_i S_i^T \right) X^T, \quad S_D = \sum_{i=1}^N \sum_{q=1, q \neq p}^c \frac{1}{N_p} \left(x_i^{(p)} - W^T n_{x_i}^{(p)} \right) \left(x_i^{(p)} - W^T n_{x_i}^{(p)} \right)^T.$$

This constrained optimization problem can be figured out by enforcing Lagrange multiplier. First, a function $J(W)$ can be linearly constructed by the objective function and the constraint:

$$\begin{aligned} J(W) &= \max \left\{ \text{tr} \left\{ W^T S_D W \right\} - \lambda \text{tr} \left(W^T S_L W - S_L \right) \right\} \\ &= \max \text{tr} \left\{ W^T S_D W - \lambda W^T S_L W + \lambda S_L \right\}. \end{aligned} \tag{19}$$

Second, the optimal projection W can be obtained from

$$\frac{\partial J(W)}{\partial W} = 2S_D W - 2\lambda S_L W = 0. \tag{20}$$

Then we have

$$S_D W = \lambda S_L W. \tag{21}$$

From Eq. (21), it can be found that the solution is composed of the eigenvectors corresponding to the d largest eigenvalues.

However, DA-LGSP often encounter the small sample size (SSS) problem when applied to real world data such as face recognition so that the matrix of locality structure is singular, since the training sample’s number is smaller than the original dimensions. To address this issue, PCA is preferred over DA-LGSP to reduce the original dimensions so that S_L is nonsingular in the PCA subspace.

3.4 Proposed DA-LGSP

The proposed DA-LGSP can be summarized as follows:

Step 1 PCA has been utilized to project the original space into a lower dimensional subspace. Denoted the transformation matrix of PCA by W_{PCA} .

Step 2 In the PCA subspace, construct the supervised KNN of every point and then use Eq. 12 to construct locality structure matrix S_L .

Step 3 Construct between-class dissimilarities S_D as Eq. 17.

Step 4 The optimal projection W is composed of the eigenvectors of $S_D W = \lambda S_L W$ corresponding to the d largest eigenvalues.

Step 5 The final projection is $W_{PCA} W$.

3.5 Computational Complexity of DA-LGSP

Assume N is the number of samples belong to c classes, N_i is the number of training samples for the i th class. D and d is the original and reduced dimensions respectively, and the number of neighbors is given by k . The computational cost of DA-LGSP includes three parts.

(1) Calculation of locality structure S_L : The first phase is the supervised k-nearest-neighbor search for which the average cost would be $O \left[\sum_1^c D \log(k) N_i \log(N_i) \right]$, and the second phase is the cost of weight matrix construction will be $O [NkD]$.

(2) Calculation of between-class dissimilarities S_D : the cost of calculating between-class dissimilarities would be $O \left[\sum_{i=1}^c \sum_{j=i}^c DN_i N_j \right]$.

(3) Eigenvalue decomposition has a cost of $O [ND^2]$.

In summary, the entire computational complexity of DA-LGSP is

$$O \left[\sum_1^c D \log(k) N_i \log(N_i) + NkD + \sum_{i=1}^c \sum_{j=i}^c DN_i N_j + ND^2 \right].$$

4 Experiment

Experiments were conducted on USPS data, COIL-20 dataset, ORL dataset and FERET dataset. We compared our proposed method with several state-of-the-art approaches for images feature extraction. The compared methods include LDA, LPP, UDP, CMVM, MMDA and NDML which are briefly introduced in Sect. 1. Most of the parameters in each method used for comparison were set according to the recommendations in the original references. In the PCA stage, we preserved nearly 95% image energy to select the number of principal components. When constructing the neighborhood graph, the KNN search was used for all methods. Moreover, the nearest-neighbors classifier is adopted to predict the labels of test data.

4.1 Experiments on USPS Database

The USPS handwriting digital data [38] include 10 classes from “0” to “9”. Each class has 1100 examples. The images in the database are manually cropped and rescaled to $16 * 16$. Figure 1 displays a subset of digital “2” from original USPS handwriting database.

Ten times experiments were repeated by randomly choosing a subset include 100 images of every class from the original database, the first l ($l = 30, 40, 50, 60, 70, 80$) images per class for training and the remaining images for testing. Each image is normalized to be a unit vector. When constructing the KNN graph, K is set to 12. The optimized average recognition rates at any possible dimensions of each method are given in Table 1.

In the second experiment, 50 images per class were randomly selected for training, and 50 images for testing. The recognition rate curves versus the variation of dimensions are shown in Fig. 2.

It is found that the proposed method outperforms other techniques with the variable of number of training or final dimensions.



Fig. 1 The sample digital images “2” from USPS handwriting database

Table 1 The optimized average recognition rates on USPS database by different methods

	LDA	LPP	UDP	CMVM	NDML	MMDA	DA-LGSP
$l = 30$	0.8187	0.7643	0.7700	0.8236	0.8234	0.8274	0.8346
$l = 40$	0.8450	0.8043	0.8120	0.8368	0.8493	0.8500	0.8740
$l = 50$	0.8660	0.8128	0.8276	0.8608	0.8552	0.8680	0.8872
$l = 60$	0.8820	0.8255	0.8313	0.8620	0.8900	0.8830	0.8990
$l = 70$	0.8760	0.8313	0.8413	0.8700	0.8807	0.8840	0.9033

Bold values indicate the best average recognition rate

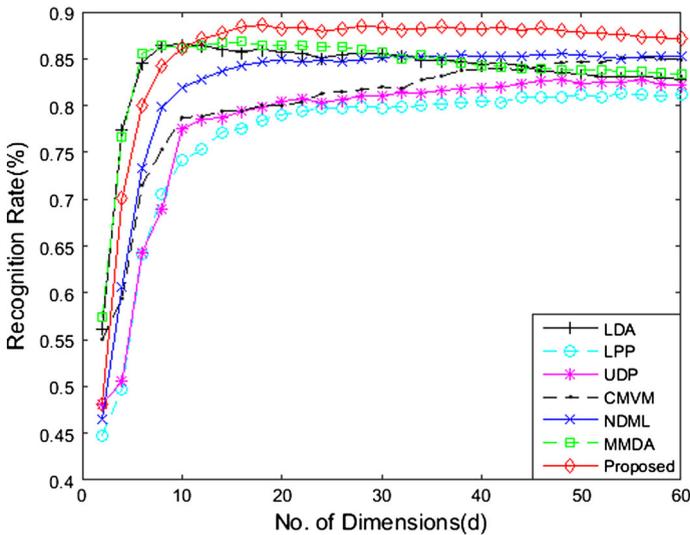


Fig. 2 The recognition rates versus different dimensions on USPS data

4.2 Experiments on COIL-20 Dataset

COIL-20 (Columbia Object Image Library) [39], a man-made object dataset consisting of 20 man-made objects, there are 72 images of different viewpoints for each object. The images are manually cropped and then normalized to 128 * 128 pixels. Samples from each class of COIL-20 dataset are shown in Fig. 3.

Firstly, we randomly select l ($l = 10, 20, 30, 40$) images per class for training and the remaining images for testing. The KNN parameter K in LPP, UDP, NDML and the proposed algorithm is chosen as 8. The maximal recognition rates of each method for all possible dimensions are given in Table 2. Secondly, the first 10 images are randomly selected as training samples and the rest 62 images as testing set. The proposed method and compared methods have been evaluated on the same training samples and the testing samples. We run the system ten times, all possible dimensions of the low-dimensional representation were evaluated, and curves of the best recognition rates versus ten different training sets are shown in Fig. 4.

As can be seen, this proposed algorithm has higher average recognition rates than others, except for only individual test compare with NDML.

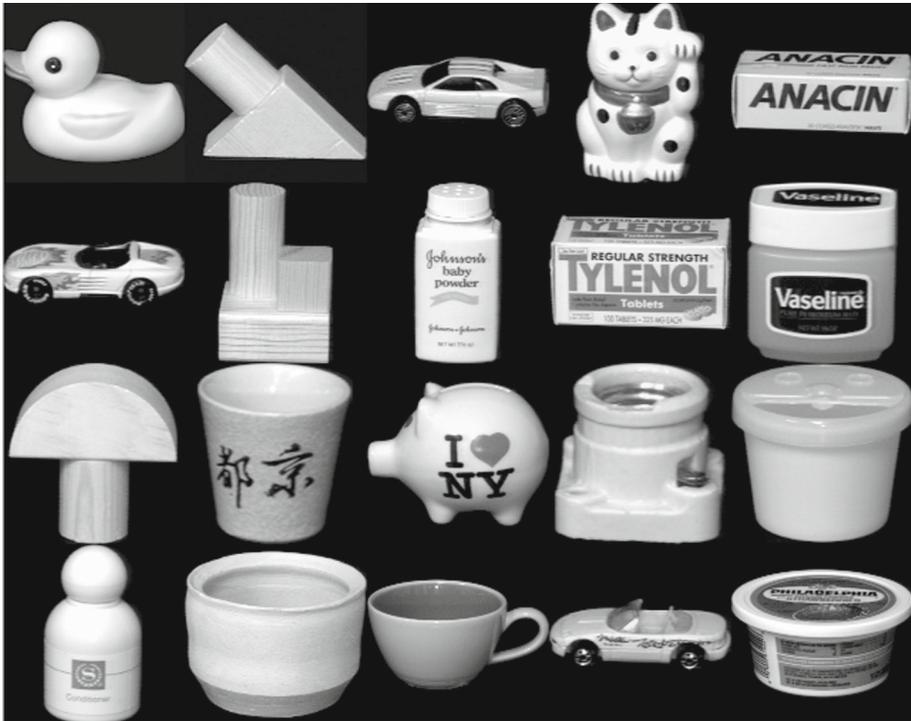


Fig. 3 Samples of 20 objects in COIL-20 dataset

Table 2 The optimized recognition rates on COIL-20 dataset by different methods

	LDA	LPP	UDP	CMVM	NDML	MMDA	DA-LGSP
$l = 10$	0.9076	0.8577	0.8495	0.8956	0.9192	0.9284	0.9313
$l = 20$	0.9588	0.9252	0.9196	0.9538	0.9804	0.9696	0.9767
$l = 30$	0.9748	0.9529	0.9543	0.9745	0.9869	0.9802	0.9900
$l = 40$	0.9791	0.9753	0.9716	0.9791	0.9931	0.9922	0.9959

Bold values indicate the best average recognition rate

4.3 Experiments on ORL Dataset

The ORL database [40] contain 400 images of 40 persons, each has ten different images with the variation of lighting conditions, facial expressions and other details. Images in the dataset are manually cropped and rescaled to $112 * 92$. Figure 5 shows a sample of ORL dataset.

In our experiments, l images (l varies from 3 to 8) are randomly selected of each individual to form the training set. The remaining $(10 - l)$ images are used for testing. The KNN parameter K in LPP, UDP, NDML and the proposed algorithm is chosen as $l - 1$. For each l , we run the system ten times. The average recognition rates of each method with the same final dimensions ($d = 30$) are given in Table 3.

And then, four images of each person are randomly chosen for training, while the remaining six images are used for testing. The parameters involved in each method are set as the

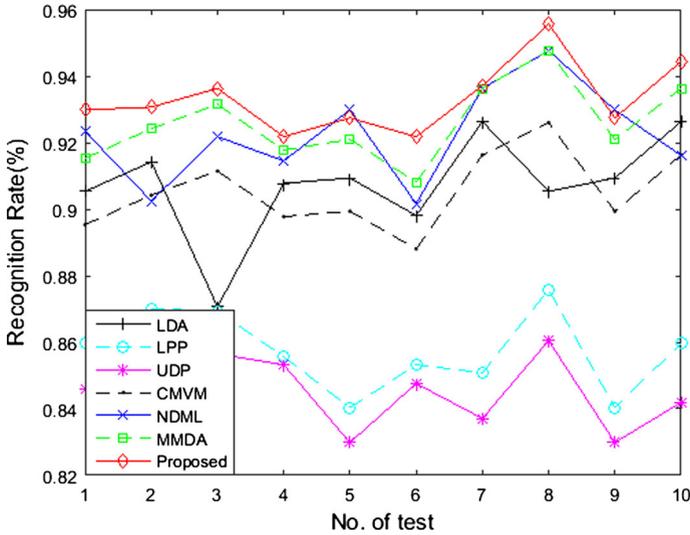


Fig. 4 The recognition rates versus different training sets on COIL-20database



Fig. 5 Ten images of one person in ORL dataset

Table 3 The average recognition rates on ORL dataset by different methods

	LDA	LPP	UDP	CMVM	NDML	MMDA	DA-LGSP
$l = 3$	0.8493	0.7221	0.7529	0.84	0.8514	0.8393	0.8607
$l = 4$	0.8675	0.7758	0.8295	0.9025	0.8983	0.9042	0.935
$l = 5$	0.906	0.806	0.805	0.927	0.927	0.939	0.941
$l = 6$	0.9462	0.8563	0.8838	0.93	0.9525	0.9513	0.9637
$l = 7$	0.9383	0.8733	0.945	0.9667	0.9683	0.9733	0.9733
$l = 8$	0.9575	0.885	0.935	0.96	0.9725	0.9725	0.975

Bold values indicate the best average recognition rate

same as those used in the first experiment. The recognition rate curves versus the variation of dimensions are shown in Fig. 6. It can be found that DA-LGSP also obtained the best classification results compared to other methods.

4.4 Experiments on FERET Dataset

The FERET dataset [41] in our experiments consists of including 1400 gray-level face images comprising 200 different people with 7 images each. There are 71 females and 129 males, who are diverse across ethnicity, gender, and age. Images in the dataset are manually cropped and rescaled to 80 * 80. Figure 7 shows images with different expressions, illuminations and poses of one person from FERET database.

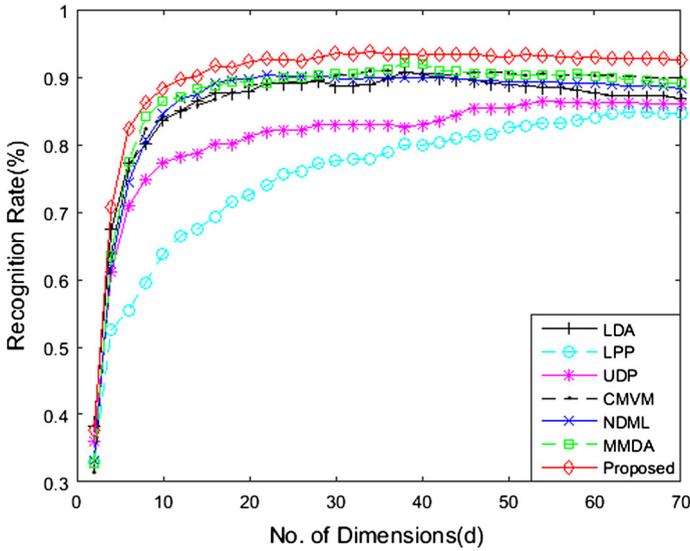


Fig. 6 The recognition rates versus different dimensions on ORL dataset



Fig. 7 Seven images of one person in FERET dataset

Table 4 The optimized average recognition rates on FERET dataset by different methods

	LDA	LPP	UDP	CMVM	NDML	MMDA	DA-LGSP
$l = 3$	0.8	0.6212	0.6198	0.796	0.7408	0.8145	0.8145
$l = 4$	0.8453	0.6635	0.6991	0.8543	0.7843	0.873	0.8793
$l = 5$	0.8795	0.7364	0.7864	0.8825	0.8065	0.897	0.9005
$l = 6$	0.884	0.8039	0.8633	0.89	0.854	0.902	0.904

Bold values indicate the best average recognition rate

As the experiments on ORL dataset, l images (l varies from 3 to 6) are randomly selected of each individual to form the training sample set. The remaining $7 - l$ images are used for testing. The KNN parameter K in LPP, UDP, NDML and the proposed algorithm is chosen as $l - 1$. Table 4 tabulates the optimal average recognition rates at any possible dimensions of different methods on FERET dataset.

Secondly, the first five images are randomly selected as training samples and the rest two images as testing set. The proposed method and compared methods have been evaluated on the same training samples and the testing samples. The parameters involved in each method are set as the same as those used in the first experiment. We run the system ten times. All possible dimensions of the low-dimensional representation were evaluated on the same training set and testing set, Fig. 8 shows the best recognition rate curves versus the ten different training sets.

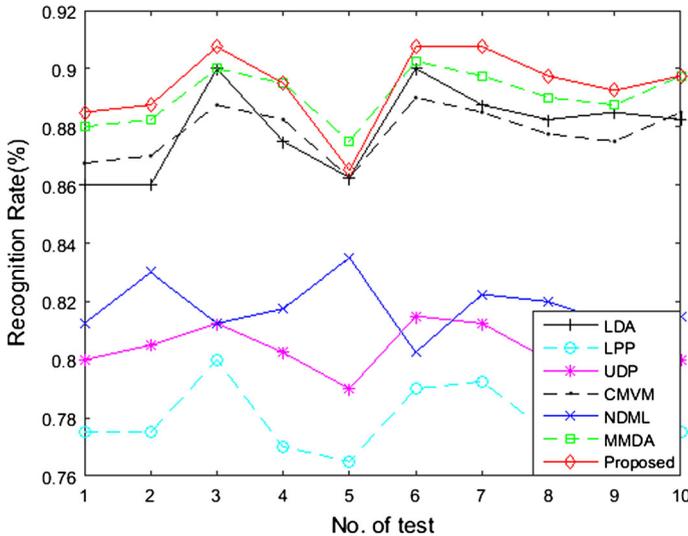


Fig. 8 The recognition rates versus different training sets on the FERET database

We can find that the effectiveness of NDML decrease apparently, while the proposed method works consistently well.

4.5 Analysis of Parameter k and $2\sigma^2$

There are three parameters in the proposed algorithm. The recognition rates versus different dimensions have been discussed in each dataset, and the effect of the parameters k and σ in DA-LGSP is analyzed in this part. The experiment is conduct on USPS database by randomly choosing subsets include 100 images of every class from the original database, and both the number of training images and testing images are fixed at 50. Firstly, σ is fixed at the average value of the distances between $x_i^{(p)}$ and its neighbors (denoted by d_{ave}), k is changed from 4 to 49 (49 is the maximum of k). Secondly, we fixed k at 12, the value of σ is changed in the range $[10^{-2}d_{ave}, 10^{-1}d_{ave}, d_{ave}, 10d_{ave}, 10^2d_{ave}, 10^3d_{ave}, 10^4d_{ave}, 10^5d_{ave}]$. Figure 9 provides the experimental results on USPS. We can observe that the performance keeps stable when the value of k changes between 4 and 20, but recognition rate decreases with the enlargement of k from 20 to 49. Besides, compared with k , the performance of σ is more stable. When σ changes between $10^{-1}d_{ave}$ to 10^2d_{ave} , the recognition rate is better than others.

Both k and σ reflect the affect of locality structure to feature extraction. Proper k can reflect well the local geometry of the manifold while k couldn't be too large, large local patch will lead to a large bias to the real embedding result. As for σ , too small or too large σ will weaken the dissimilarities of the neighbors to every data point.

4.6 Discussion

According to the experiments being systematically performed on the four datasets, we can find that:

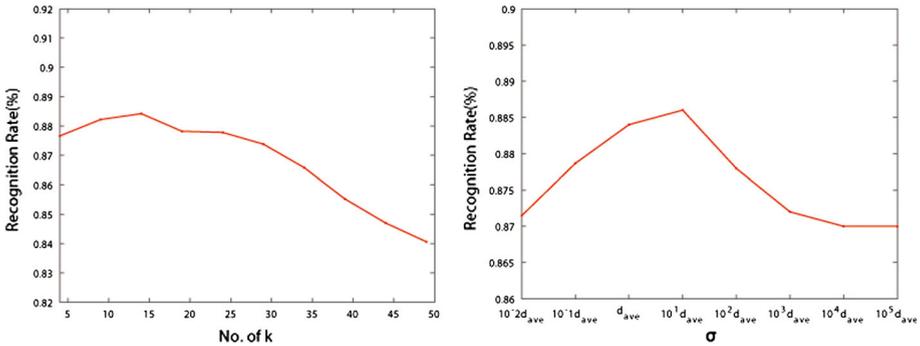


Fig. 9 Average recognition rates with different parameter configurations

(1) It can be found that the recognition rates show the increasing trend with the increasing of dimensions. However, when the recognition rates achieve its maxima, they almost keep unchanged.

(2) Frankly speaking, the supervised methods such as DA-LGSP, MMDA, NDML, LDA and CMVM perform better than the unsupervised ones. The proposed method is a supervised one based on manifold learning, it can gain the best recognition rates among the methods involved in the experiments.

(3) All of the supervised methods mentioned in this section include the proposed methods are linear subspace learning methods based on Fisher framework, compared with LDA, MMDA, NDML and CMVM, DA-LGSP considers both the local topology structure and the between-class dissimilarities. Notwithstanding NDML and MMDA perform well in some dataset, experimental results on handwriting digital data, man-made objects and two face datasets verifies that the proposed method has the best performance on all of the datasets and surpasses other competing methods.

(4) Under the analysis of parameter k and σ , we suggest that the value of k would not be larger than 20, and choose σ as the average value of the distances between $x_i^{(p)}$ and its neighbors.

5 Conclusions

In this paper, we have proposed a feature extraction method based on manifold learning which has fully considered class labels and local geometry. To avoid the out-of-sample problem, DA-LGSP focuses on developing a linear transformation that make different classes separated as much as possible in the final embedding space under the constraint of local preserving. It has shown that the linear transformation can maximize the dissimilarities between all the classes. As a result, it leads to stable and reasonable recognition rates of testing sample. Our proposed method achieves the best performances comparing with several state-of-the-art methods on four commonly used image datasets.

References

1. Dalton L, Saurabh P, Melba MC, Okan E (2014) Manifold-learning-based feature extraction for classification of hyperspectral data: a review of advances in manifold learning. *IEEE Signal Proc Mag* 31(1):55–66

2. He X, Yan S, Hu Y, Niyogi P (2005) Face recognition using laplacianfaces. *IEEE Trans Pattern Anal Mach Intell* 27(3):328–340
3. Yu J, Tao D, Wang M, Rui Y (2015) Learning to rank using user clicks and visual features for image retrieval. *IEEE Trans on Cybern* 45(4):767–779
4. Yu J, Tao D, Wang M (2012) Adaptive hypergraph learning and its application in image classification. *IEEE Trans Image Process* 21(7):3262–3272
5. Yu J, Rui Y, Tao D (2014) Click prediction for web image reranking using multimodal sparse coding. *IEEE Trans Image Process* 23(5):2019–2032
6. Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 15(6):1299–1319
7. Liu T, Tao D, Song M, Maybank S (2017) Algorithm-dependent generalization bounds for multi-task learning. *IEEE Trans Pattern Anal Mach Intell* 39(2):227–241
8. Liu T, Tao D (2015) On the performance of manhattan nonnegative matrix factorization. *IEEE Trans Neural Netw Learn Syst* 27(9):1851–1863
9. Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791
10. Liu T, Tao D, Xu D (2016) Dimensionality-dependent generalization bounds for k-dimensional coding schemes. *Neural Comput* 28(10):2213
11. Liu T, Gong M, Tao D (2016) Large-cone nonnegative matrix factorization. *IEEE Trans Neural Netw Learn Syst*. doi:[10.1109/TNNLS.2016.2574748](https://doi.org/10.1109/TNNLS.2016.2574748)
12. Sam R, Lawrence S (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326
13. Tenenbaum J, Silva V, Langford J (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323
14. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15(6):1373–1396
15. Zhang Z, Zha H (2004) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J Sci Comput* 26(1):313–338
16. Donoho D, Grimes C (2003) Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc Natl Acad Arts Sci* 100(10):5591–5596
17. Chen J, Ma Z, Liu Y (2013) Local coordinates alignment with global preservation for dimensionality reduction. *IEEE Trans Neural Netw Learn Syst* 24(1):106–117
18. Niu G, Ma Z, Lv S (2017) Ensemble multiple-kernel based manifold regularization. *Neural Process Lett* 45:539–552
19. Dollar P, Rabaud V, Belongie S (2007) Non-isometric manifold learning: analysis and an algorithm. In: *Proceedings of international conference on machine learning*, pp 241–248
20. Farahmand A, Szepesvari C, Audibert J (2007) Manifold-adaptive dimension estimation. In: *Proceedings of international conference on Machine learning*, pp 265–272
21. Gerber S, Tasdizen T, Whitaker R (2007) Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian eigenmaps. In: *Proceedings of international conference on machine learning*, pp 281–288
22. Huang D, Yi Z, Pu X (2009) Manifold-based learning and synthesis. *IEEE Trans Syst Man Cybern B Cybern* 39(3):592–606
23. Mordohai P, Medioni G (2010) Dimensionality estimation, manifold learning and function approximation using tensor voting. *J Mach Learn Res* 11(1):411–450
24. Zhang C, Xiang S, Nie F, Song Y (2009) Nonlinear dimensionality reduction with relative distance comparison. *Neurocomputing* 72(7–9):1719–1731
25. Yang Y, Nie F, Xiang S, Zhuang Y, Wang W (2010) Local and global regressive mapping for manifold learning with out-of-sample extrapolation. *Proc AAAI Conf Artif Intell* 1:649–654
26. Xiang S, Nie F, Pan C, Zhang C (2011) Regression reformulations of LLE and LTSA with locally linear transformation. *IEEE Trans Syst Man Cybern B* 41(5):1250–1262
27. Xiang S, Nie F, Zhang C, Zhang C (2009) Nonlinear dimensionality reduction with local spline embedding. *IEEE Trans Knowl Data Eng* 21(9):1285–1298
28. He X, Niyogi P (2005) Locality preserving projections. *Adv Neural Inf Process Syst* 16(1):186–197
29. Yang J, Zhang D, Yang J, Niu B (2007) Globally maximizing, locally minimizing: unsupervised discriminant projection with application to face and palm biometrics. *IEEE Trans Pattern Anal* 29(4):650–664
30. Yu J, Wang M, Tao D (2012) Semi-supervised multiview distance metric learning for cartoon synthesis. *IEEE Trans Image Process* 21(11):4636–4648
31. Xu C, Tao D, Rui Y (2014) Large-margin weakly supervised dimensionality reduction. *International Conference on Machine Learning, ICML 3:2472–2482*

32. Yu J, Yang X, Gao F, Tao D (2016) Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Trans on Cybernetics*. doi:[10.1109/TCYB.2016.2591583](https://doi.org/10.1109/TCYB.2016.2591583)
33. Li B, Huang D, Wang C, Liu K (2008) Feature extraction using constrained maximum variance mapping. *Pattern Recognit* 41(11):287–3294
34. Yang W, Sun C, Zhang L (2011) A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognit* 44(8):1649–1657
35. Bian W, Tao D (2011) Max-min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE Trans Pattern Anal Mach Intell* 33(5):1037–1050
36. Lu J, Tan Y, Wang G (2013) Discriminative multimanifold analysis for face recognition from a single training sample per person. *IEEE Trans Pattern Anal* 35(1):39–51
37. Li B, Li J, Zhang X (2015) Nonparametric discriminant multi-manifold learning for dimensionality reduction. *Neurocomputing* 152(25):121–126
38. RnavGraphImageData. USPS Handwritten Digits. <http://www.cs.toronto.edu/~roweis/data.html>
39. Nene S, Nayar S, Murase H, et al. (1996) Columbia object image library (coil-20), Technical Report CUCS-005-96
40. Olivetti & Oracle Research Laboratory (1994) The Olivetti & Oracle Research Laboratory face database of faces. <http://www.cam-orl.co.uk/facedatabase.html>
41. Phillips P (2006) The Facial Recognition Technology (FERET) database. http://www.itl.nist.gov/iad/humanid/feret/feret_master.html