*ICIC Express Letters*
*Part B: Applications*
Volume **8**, Number **1**, January **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 11–18

# CLUSTERING-BASED DISCRIMINATIVE SPARSITY PRESERVING PROJECTIONS FOR UNSUPERVISED FACE RECOGNITION

Yongxin Wang and Huaxiang Zhang*

Department of Computer Science
Shandong Normal University
No. 88, East Wenhua Road, Lixia District, Jinan 250014, P. R. China
*Corresponding author: wangyongxin1992@126.com

Received July 2016; accepted October 2016

Abstract. *Recently, dimension reduction algorithms are widely applied to high-dimensional data preprocessing, especially for face images. In this paper, a novel unsupervised dimension reduction algorithm, named clustering-based discriminative sparsity preserving projections (CDSPP), is proposed by integrating cluster analysis and sparse representation analysis into a joint framework. Unlike many existing approaches such as sparsity preserving projections (SPP), where the constructive weights are computed by the classical sparse representation (SR), CDSPP introduces some class discriminant information by using clustering firstly, and CDSPP uses margin maximization criterion (MMC) to apply the discriminative information in the learning model. The obtained projections will contain more discriminant information than classical sparse subspace learning methods SPP. Moreover, CDSPP is an unsupervised dimensionality reduction method, which improves the simplicity of model training. Experiments on AR and Yale-B image datasets demonstrate its effectiveness.*

**Keywords:** Dimensionality reduction, Clustering-based discriminative sparse representation, Sparse subspace learning, Face recognition

1. **Introduction.** The dimensionality of variables or feature is usually very high in many real world application domains, such as face recognition, signal processing, and text categorization. These high-dimensional features may bring some disadvantages, such as over-fitting, low efficiency and poor performance. To mitigate the so-called “curse of dimensionality” [1] and to improve the computational efficiency, dimensionality reduction (DR) is an effective approach to preprocessing such data. So far, a variety of dimensionality reduction methods for projecting the high-dimensional data from their original space into low-dimensional feature spaces have been proposed and studied for machine learning applications. These traditional methods can be categorized as supervised algorithms, semi-supervised algorithms and unsupervised algorithms, based on whether the sample labels are considered for dimensionality reduction [2]. Supervised approaches use the discriminative information encoded in the labels to learn a more discriminative subspace. Some supervised sparsity-based approaches have been studied [3, 4, 5] to improve the discrimination ability of learning models. However, the label information is hard to obtain and a data set usually has small labeled data and large unlabeled data. So the so-called “small labeled sample problem” [6] is usually a challenge for supervised algorithms. Due to the small labeled data, semi-supervised algorithms [7, 8, 9] are developed to exploit both the labeled and unlabeled data simultaneously. In many real world applications, data is usually high-dimensional and with no label. Therefore, it is quite necessary and promising to develop unsupervised dimensionality reduction algorithms [10]. So in this paper, we only focus on unsupervised methods, due to its effectiveness and feasibility.

In the unsupervised DRs, principle component analysis (PCA) [11] seems to be the most popular one. It aims to keep the variance of data as much as possible, and the reduced dimensions are linear combinations of original features. Besides, enormous manifold-based methods, which are based on the idea that data are usually samples from a low-dimensional manifold that is embedded in a high-dimensional space, such as local linear embedding (LLE) [12] and Laplacian eigenmaps (LE) [13] have been developed to explicitly discover the nonlinear manifold structure concealed in the data. In LLE and LE, it is assumed that each data point can be linearly reconstructed by its nearest neighbors, and in a low-dimensional space, its representation should also be linearly reconstructed by the representations of its nearest neighbors with the same reconstruction coefficients. However, they have the out of sample problem [14]. Thus locality preserving projections (LPP) [15] and neighborhood preserving embedding (NPE) [16] were proposed to solve this problem. In fact, LPP and NPE can be seen as linear models of LE and LLE respectively.

Recently, some new approaches integrating the theory of sparse representation have been proposed and have been successfully applied in many real world applications. Sparse subspace learning (SSL) [17] is a special kind of dimensionality reduction methods which considers "sparsity". It finds a subspace spanned by sparse base vectors and the sparsity is embedded on the projection vectors. Qiao et al. [18] firstly presented a sparsity preserving projection (SPP), which aims to preserve the sparse reconstructive relationship of the data by minimizing an L1-objective function. All these algorithms are designed based on classical SR. However, the classical unsupervised SR is a global linear method which tends to lose the local information of data. It has been shown that locality is more essential than sparsity in some case [19]. Thus in order to solve the problem in classical unsupervised SR, Li et al. [20] proposed a clustering-guided sparse structural learning for unsupervised feature selection (CGSSL) by integrating cluster analysis and sparse structural analysis into a joint framework.

Motivated by the recent development of SR, we propose a novel unsupervised dimensionality reduction algorithm, namely clustering-based discriminative sparsity preserving projections (CDSPP), which integrates cluster analysis and sparse representation analysis into a joint framework. CDSPP uses a model in which the cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities to cluster each point of the training data and divides the date into several clusters firstly. Then margin maximization criterion (MMC) is used to encode discriminant label information. Finally a classical sparse representation (SR) is used. Compared to SPP, the coefficients computed by CDSPP are more discriminative and take more local structure of data into account. Extensive experiments are conducted on human face images. The experimental results show that the proposed algorithm is superior to several relative algorithms.

The rest of the paper is organized as follows. In Section 2, we review the related work of SPP and give out the overview of our proposed method. Experimental results and comparisons on two real-world datasets are demonstrated in Section 3. Finally, the conclusion is given in Section 4.

## 2. Clustering-Based Discriminative Sparsity Preserving Projections.

### 2.1. Sparsity preserving projections (SPP).

Sparse representation (SR) has been successfully applied to solve several real world problems such as pattern recognition and machine learning. SR aims to represent each data point $\mathbf{x}_i$ using as few entries of $\mathbf{X}$ as possible. SPP first computes a sparse reconstructive weight vector $\mathbf{s}_i$ for each data point $\mathbf{x}_i$ by solving the following minimization problem [18]:

$$\hat{\mathbf{s}}_i = \arg\min_{||\mathbf{s}_i||=1} ||\mathbf{s}_i||_1, \text{ s.t. } \mathbf{x}_i = \tilde{\mathbf{X}}\mathbf{s}_i \tag{1}$$

where $\mathbf{s}_i = [s_{i1}, \ldots, s_{i,i-1}, 0, s_{i,i+1}, \ldots, s_{in}]^T$. In fact, $\tilde{\mathbf{X}}$ is a subset of all the given data $\mathbf{X}$ which does not include $\mathbf{x}_i$, so the $i^{\text{th}}$ element of $\mathbf{s}_i$ should be zero. Then we construct the sparse reconstructive weight matrix $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n]$.

A reasonable criterion for seeking a "good" projection $\beta$ which best preserves the optimal weight vector $\hat{\mathbf{s}}_i$ is to minimize the following cost function [18]:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left|\left|\beta^T \mathbf{x}_i - \beta^T \mathbf{X}\hat{\mathbf{s}}_i\right|\right|^2 = \arg\min_{\beta} \beta^T \mathbf{X}(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{X}^T \beta \tag{2}$$

Then the optimization problem of (2) can be solved by calculating the generalized eigenvectors of the following generalized eigenvector problem:

$$\mathbf{X}(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{X}^T \beta = \lambda \mathbf{X}\mathbf{X}^T \beta \tag{3}$$

2.2. **Margin maximization criterion (MMC).** Margin maximization criterion (MMC) [21] is an existing dimensionality reduction algorithm. Its main idea is to maximize the margin between classes. We can derive linear discriminant analysis (LDA) from MMC by adding some suitable constraints. And MMC overcomes the drawback of small sample size problem that LDA has.

The objective function of MMC can be written as follows

$$J = \frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} \left[p_i p_j (d(\mathbf{m}_i - \mathbf{m}_j) - s(\mathbf{m}_i) - s(\mathbf{m}_j))\right] \tag{4}$$

where $c$ is the number of classes, $p_i$ and $p_j$ are the prior probability of class $i$ and class $j$, and $\mathbf{m}_i$ and $\mathbf{m}_j$ are the mean vectors of class $i$ and class $j$. $d(\mathbf{m}_i - \mathbf{m}_j)$, $s(\mathbf{m}_i)$, and $s(\mathbf{m}_j)$ are defined as the following respectively.

$$d(\mathbf{m}_i - \mathbf{m}_j) = ||\mathbf{m}_i - \mathbf{m}_j||, \quad s(\mathbf{m}_i) = tr(\mathbf{C}_i), \quad s(\mathbf{m}_j) = tr(\mathbf{C}_j) \tag{5}$$

where $\mathbf{C}_i$ is the covariance matrix of class $i$.

It is easy to show that optimization (5) can be reformulated as

$$J = tr(\mathbf{S}_b - \mathbf{S}_w) \tag{6}$$

where

$$\mathbf{S}_b = \sum_{i=1}^{c} p_i \left(\mathbf{m}_i - \sum_{j=1}^{c} p_j \mathbf{m}_j\right) \left(\mathbf{m}_i - \sum_{j=1}^{c} p_j \mathbf{m}_j\right)^T \tag{7}$$

is called the between-class scatter matrix and

$$\mathbf{S}_w = \sum_{i=1}^{c} p_i \mathbf{C}_i \tag{8}$$

is called the within-class scatter matrix.

2.3. **The CDSPP algorithm.** SPP is effective in many domains, but it is unsupervised and its unsupervised nature restricts its discriminating capability. Even though some discriminative SPP [3, 4, 5] approaches are proposed to improve the discriminant property of SPP, they are easy to get into the trouble of "small labeled sample problem". Motivated by this reason, we propose a novel approach called CDSPP which is unsupervised and discriminative.

We use the classical sparse representation as introduced in Section 2.1.

$$\hat{\mathbf{s}}_i = \arg \min_{||\mathbf{s}_i||=1} ||\mathbf{s}_i||_1, \text{ s.t. } \mathbf{x}_i = \tilde{\mathbf{X}}\mathbf{s}_i \tag{9}$$

And its optimized function can be derived into the following

$$J_1 = \min_{\beta} \sum_{i=1}^{n} \left|\left|\beta^T \mathbf{x}_i - \beta^T \mathbf{X}\hat{\mathbf{s}}_i\right|\right|^2 = \min_{\beta} \beta^T \mathbf{X}(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{X}^T \beta \tag{10}$$

Furthermore, the main idea of linear transformation is to find a mapping $\beta$ to project a new pattern further from patterns in different classes but closer to those in the same class, which is exactly the goal of classification. That is to say, MMC is a good criterion being used in classification problems. However, MMC is a supervised algorithm and label information is used in the algorithm. So we should obtain the discriminative information of unlabeled data. We firstly use clustering algorithm to divide data into several parts.

For each data point $\mathbf{x}_i$, we compute two variables: its local density $\rho_i$ and its distance $\sigma_i$ from points of higher density [22]. Both the variables are computed depending only on the distance $d_{ij}$ between data points. The local density $\rho_i$ is defined as

$$\rho_i = \sum_{j=1}^{n} \varphi(d_{ij} - d_c) \tag{11}$$

where $\varphi$ is a function. $\varphi(d_{ij} - d_c)$ equals 1 when $d_{ij} - d_c < 0$ and 0 otherwise and $d_c$ is a cutoff distance. $\rho_i$ is equivalent to the number of points within a circle, the center and radius of which are $\mathbf{x}_i$ and $d_c$ respectively. Besides, $\sigma_i$ is defined as follows

$$\sigma_i = \min d_{ij}, \ \ j : \rho_j > \rho_i \tag{12}$$

$\sigma_i$ is a relative distance which is the smallest value of distances between $\mathbf{x}_i$ and other with higher density points. For the point with highest density, we define $\sigma_i = \max d_{ij}$. Obviously, the clusters are those which have high density and are far from other clusters. If $\mathbf{x}_i$ has a big $\sigma_i$ and small $\rho_i$, $\mathbf{x}_i$ is a noise point probably. What is more, if $\mathbf{x}_i$ has a big $\rho_i$ and small $\sigma_i$, $\mathbf{x}_i$ is very likely a point besides cluster. So we define the following clustering criterion:

$$\varepsilon_i = \rho_i \sigma_i \tag{13}$$

We sort $\varepsilon$ in descending order, and select those points corresponding to largest values as the cluster centers. Then for each given data $\mathbf{x}_i$, we classify it to the cluster which $\mathbf{x}_i$ is closest to. Finally, all given data are clustered into $c$ parts.

Then we use MMC to make the distance between clusters as large as possible while the distance within a cluster as small as possible. When performing dimensionality reduction, we want to find a mapping from the original data to some feature subspace in which $J$ is maximized after the transformation. We suppose there is a linear transformation $\mathbf{Y} = \beta^T \mathbf{X}$ that maximizes $J$, and an optimal subspace will be derived as follows

$$J_2 = \max_{\beta} tr \left[\beta^T (\mathbf{S}_b - \mathbf{S}_w) \beta\right] \tag{14}$$

Obviously, if the transformation $J_1$ obtained by SPP can solve $J_2$ simultaneously, the discriminative information will be imposed greatly. Thus the solution of CDSPP can be derived into the following optimization problem:

$$\hat{\beta} = \arg \min_{\beta^T \mathbf{X}\mathbf{X}^T \beta = 1} tr \left\{\beta^T \left[\mathbf{X}(\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S}) \mathbf{X}^T - \mu(\mathbf{S}_b - \mathbf{S}_w)\right] \beta\right\} \tag{15}$$

where $\mu$ is a parameter to balance the sparsity and the discriminative information.

Then the optimal $\hat{\beta}$ of Formula (16) is given by the minimum eigenvalue solution to the following generalized eigenvector problem:

$$\left[\mathbf{X}(\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S}) \mathbf{X}^T - \mu(\mathbf{S}_b - \mathbf{S}_w)\right] \beta = \lambda \mathbf{X}\mathbf{X}^T \beta \tag{16}$$

Thus, the projection can be written as follows

$$\mathbf{Y} = \hat{\beta}^T \mathbf{X} \tag{17}$$

3. **Experiment.** In this section, we have conducted two experiments on the popular face databases AR and Yale-B to verify the effectiveness of the proposed CDSPP method. We compare CDSPP with several typical dimensionality reduction methods such as PCA, LPP, NPE and SPP. A 1-NN classifier is employed to classify the projected feature space. To robustly evaluate the performance of different algorithms under different sample condition, we use 5-fold cross validation and all the experiments are implemented on MATLAB platform. All images in Yale-B are resized to $32 \times 32$ pixels, and images in AR are resized to $60 \times 43$. We adopt the implementation in SPAMS package to solve Formula (9) and $d_c$ was set to satisfy the average number of $\rho$ is around 0.02 of the total number of points in the dataset. We set 1 to $\mu$ in CDSPP empirically.

3.1. **The AR face image database.** The AR database consists of over 4000 frontal face images of 126 individuals with different facial expressions, occlusions and lighting conditions. Figure 1 gives some samples of this dataset. We chose a subset of the dataset consisting of 50 female and 50 male subjects.

Table 1 shows the recognition rates and their corresponding standard deviations of each method under ten different dimensions when using 1-NN classifier. The rank of each method is shown on the right of the recognition rate. The best algorithm ranks one, and the worst ranks five. If both algorithms perform the same, average ranks are assigned [23]. An average rank is computed for each algorithm. We also perform t-test on the compared methods, and the results are shown in Table 2. Each column shows p-values of t-test on CDSPP and other methods under different feature dimensionality. The p-values indicate the probability that two sample sets are distributed with equal means. The smaller the value is, the bigger difference the two algorithms have, and 0.05 is a threshold.

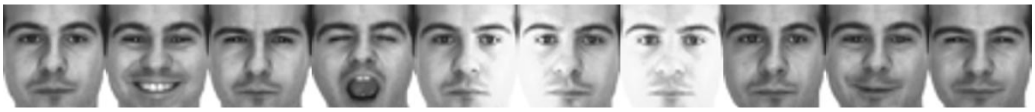

Figure 1. Sample face images from AR database

Table 1. Recognition rate on AR database under different dimensionalities

| | 20 | | 40 | | 60 | | 80 | | 100 | | 120 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PCA | 65.73±2.53 | 2 | 73.79±2.48 | 3 | 75.64±1.07 | 3 | 76.93±0.99 | 3 | 77.80±1.35 | 3 | 78.23±1.54 | 3 |
| LPP | 65.73±1.27 | 5 | 18.72±2.51 | 5 | 27.43±1.54 | 5 | 30.93±1.58 | 5 | 36.21±1.66 | 5 | 39.71±0.69 | 5 |
| NPE | 11.31±4.19 | 4 | 23.60±3.62 | 4 | 31.00±3.98 | 4 | 36.51±4.30 | 4 | 39.79±4.24 | 4 | 44.65±2.21 | 4 |
| SPP | 64.86±1.13 | 3 | 75.16±3.31 | 2 | 77.31±3.36 | 2 | 78.95±2.97 | 2 | 80.16±2.16 | 2 | 80.57±2.75 | 2 |
| CDSPP | **71.36±2.86** | 1 | **87.13±1.92** | 1 | **91.63±1.20** | 1 | **92.63±1.10** | 1 | **91.99±1.33** | 1 | **92.35±1.27** | 1 |
| | 140 | | 160 | | 180 | | 200 | | Score | | | |
| PCA | 78.44±1.56 | 3 | 78.51±1.29 | 3 | 78.72±1.36 | 3 | 78.65±1.55 | 3 | 2.9 | | | |
| LPP | 40.64±0.62 | 5 | 42.93±1.68 | 5 | 44.36±0.38 | 5 | 43.65±1.17 | 5 | 5 | | | |
| NPE | 45.86±1.96 | 4 | 47.28±3.31 | 4 | 48.71±3.87 | 4 | 48.43±3.48 | 4 | 4 | | | |
| SPP | 80.56±2.81 | 2 | 80.29±2.63 | 2 | 80.67±5.89 | 2 | 80.97±4.93 | 2 | 2.1 | | | |
| CDSPP | **92.86±0.40** | 1 | **92.92±0.94** | 1 | **92.49±1.18** | 1 | **91.49±1.51** | 1 | 1 | | | |

Table 2. P-values of t-test on AR database under different dimensionalities

| | 20 | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 180 | 200 |
|---|---|---|---|---|---|---|---|---|---|---|
| CDSPP/PCA | **0.0136** | **0.0006** | **2.7E-06** | **4.1E-05** | **0.0001** | **0.0001** | **0.0001** | **0.0001** | **0.0003** | **0.0008** |
| CDSPP/LPP | **7.8E-07** | **3.6E-06** | **1.0E-06** | **9.6E-07** | **4.8E-07** | **7.5E-08** | **4.3E-09** | **1.5E-06** | **2.0E-07** | **8.4E-07** |
| CDSPP/NPE | **0.0001** | **1.4E-05** | **1.3E-05** | **2.0E-05** | **4.2E-05** | **5.7E-06** | **1.5E-06** | **1.4E-05** | **1.6E-05** | **8.1E-06** |
| CDSPP/SPP | **0.0165** | **0.0079** | **0.0028** | **0.0021** | **0.0018** | **0.0006** | **0.0008** | **0.0010** | **0.0092** | **0.0036** |

The results of Table 1 demonstrate that, CDSPP outperforms all other methods under all dimensions on recognition rate. Table 2 shows that, the differences between CDSPP and PCA, LPP, NPE and SPP are obvious under almost all dimensions. That is to say, CDSPP is distinctly superior to other methods in all dimensions.

3.2. **The Yale-B face image database.** The Yale-B database consists of 2414 frontal face images of 38 individuals under various lighting conditions. Figure 2 gives some samples of this dataset.

The recognition rates of five different methods under ten different dimensions, the corresponding dimensionality and the standard deviations of 5-fold cross validation are shown in Table 3. We also compute an averaged rank for each algorithm to evaluate their performance. The best results are highlighted in bold face. In addition, Table 4 shows the p-values of t-test of CDSPP compared with the other four methods by ten dimensions. Those values under threshold are highlighted in bold face, which show that CDSPP is significantly different from the others.

Table 3 shows that CDSPP obtains the best averaged rank and it indicates that CDSPP achieves the best recognition rates of all five different methods. When the dimension is 20 and 40, CDSPP loses to SPP or NPE according to the recognition rate, but the p-values shown in Table 4 indicate that both algorithms have no obvious difference. Based on the recognition rates shown in Table 3 and the p-values shown in Table 4, we can draw such a conclusion that CDSPP outperforms SPP when the dimension is equal or larger than 60, and there is no significant difference between CDSPP and SPP when dimension is less than 60.

Based on the results on AR and Yale-B, we may draw such a conclusion that CDSPP achieves a good performance on face recognition. CDSPP uses distance to divide data

FIGURE 2. Sample face images from Yale-B database

TABLE 3. Recognition rate on Yale-B database under different dimensionalities

| | 20 | | 40 | | 60 | | 80 | | 100 | | 120 | | 140 | | 160 | | 180 | | 200 | | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PCA | 37.87±1.40 | 4 | 53.31±0.51 | 4 | 61.18±0.42 | 4 | 64.99±1.17 | 4 | 68.39±1.27 | 4 | 69.80±1.2 | 4 | 70.92±1.16 | 4 | 71.79±0.98 | 4 | 72.45±1.23 | 4 | 72.83±1.03 | 4 | 4 |
| LPP | 32.27±0.96 | 5 | 45.16±0.93 | 5 | 54.35±0.87 | 5 | 60.52±0.59 | 5 | 64.95±2.02 | 5 | 66.86±2.01 | 5 | 69.72±2.00 | 5 | 71.05±1.87 | 5 | 71.96±1.62 | 5 | 72.75±1.80 | 5 | 5 |
| NPE | **68.89±0.50** | 1 | 77.22±1.10 | 3 | 79.91±1.03 | 3 | 80.91±0.86 | 3 | 81.73±0.47 | 3 | 82.65±0.97 | 3 | 83.10±0.98 | 3 | 83.47±0.97 | 3 | 83.27±1.14 | 3 | 83.85±1.24 | 3 | 2.8 |
| SPP | 68.23±1.15 | 3 | **83.02±0.69** | 1 | 85.80±1.37 | 2 | 87.37±1.42 | 2 | 88.12±1.35 | 2 | 88.24±1.02 | 2 | 88.65±0.78 | 2 | 89.36±0.86 | 2 | 89.94±0.87 | 2 | 89.36±0.65 | 2 | 2 |
| CDSPP | 68.27±1.33 | 2 | 82.98±0.76 | 2 | **86.04±1.05** | 1 | **88.08±1.37** | 1 | **89.32±0.81** | 1 | **89.94±0.99** | 1 | **90.52±1.14** | 1 | **90.56±1.23** | 1 | **90.68±1.23** | 1 | **90.31±0.99** | 1 | 1.2 |

TABLE 4. P-values of t-test on Yale-B database under different dimensionalities

| | 20 | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 180 | 200 |
|---|---|---|---|---|---|---|---|---|---|---|
| CDSPP/PCA | **1.3E-05** | **1.9E-07** | **3.1E-06** | **4.0E-05** | **2.1E-05** | **3.3E-05** | **4.0E-05** | **3.7E-05** | **0.0001** | **4.6E-05** |
| CDSPP/LPP | **2.5E-06** | **3.5E-07** | **5.6E-07** | **1.1E-06** | **3.3E-05** | **6.1E-06** | **4.2E-06** | **1.4E-05** | **3.6E-06** | **2.7E-06** |
| CDSPP/NPE | 0.4293 | **2.4E-05** | **0.0001** | **1.2E-05** | **7.4E-06** | **0.0001** | **0.0001** | **1.7E-05** | **1.1E-05** | **3.4E-06** |
| CDSPP/SPP | 0.8484 | 0.7091 | 0.3271 | 0.0716 | **0.0179** | **0.0002** | **0.0123** | **0.0142** | 0.2904 | **0.0112** |

into several parts and puts this discriminative information into MMC, which improves the discriminative ability of our algorithm. So it outperforms SPP on data that local structure is essential for discrimination.

4. **Conclusion.** In this paper, we proposed a clustering-based discriminative sparsity preserving projections (CDSPP). The CDSPP algorithm firstly solves a clustering problem, and then combines SPP and MMC methods. Experiments on AR and Yale-B database show the effectiveness of our method. However, the proposed approach still has room for improvement. For instance, other proposed clustering methods with a reduced time complexity could be introduced to improve the computational efficiency. In a principled manner, it remains an important direction for future work.

**Acknowledgement.** The work is partially supported by the National Natural Science Foundation of China (Nos. 61572298, 61373081, 61402268, 61401260) and the Taishan Scholar Project of Shandong, China.

## REFERENCES

[1] A. K. Jain, R. P. W. Duin and J. Mao, Statistical pattern recognition: A review, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.22, no.1, pp.4-37, 2000.

[2] H. Liu and L. Yu, Toward integrating feature selection algorithms for classification and clustering, *IEEE Trans. Knowledge and Data Engineering*, vol.17, no.4, pp.491-502, 2005.

[3] Q. Gao, Y. Huang, H. Zhang et al., Discriminative sparsity preserving projections for image recognition, *Pattern Recognition*, vol.48, no.8, pp.2543-2553, 2015.

[4] J. Gui, Z. Sun, W. Jia et al., Discriminant sparse neighborhood preserving embedding for face recognition, *Pattern Recognition*, vol.45, no.8, pp.2884-2893, 2012.

[5] G. F. Lu, Z. Jin and J. Zou, Face recognition using discriminant sparsity neighborhood preserving embedding, *Knowledge-Based Systems*, vol.31, pp.119-127, 2012.

[6] A. Jain and D. Zongker, Feature selection: Evaluation, application, and small sample performance, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.19, no.2, pp.153-158, 1997.

[7] H. Cheng, Z. Liu and J. Yang, Sparsity induced similarity measure for label propagation, *The 12th IEEE International Conference on Computer Vision*, pp.317-324, 2009.

[8] F. Zang and J. S. Zhang, Label propagation through sparse neighborhood and its applications, *Neurocomputing*, vol.97, pp.267-277, 2012.

[9] Y. Peng, B. L. Lu and S. Wang, Enhanced low-rank representation via sparse manifold adaption for semi-supervised learning, *Neural Networks*, vol.65, pp.1-17, 2015.

[10] P. Mitra, C. A. Murthy and S. K. Pal, Unsupervised feature selection using feature similarity, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.24, no.3, pp.301-312, 2002.

[11] I. T. Jolliffe, *Principal Component Analysis (Springer Series in Statistics)*, Springer-Verlag, New York, 1986.

[12] S. T. Roweis and L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, vol.290, no.5500, pp.2323-2326, 2000.

[13] M. Belkin and P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, *Neural Information Processing Systems*, vol.14, pp.585-591, 2001.

[14] Y. Bengio, J. F. Paiement, P. Vincent et al., Out-of-sample extensions for LLE, IsoMap, MDS, eigenmaps, and spectral clustering, *Advances in Neural Information Processing Systems*, vol.16, pp.177-184, 2004.

[15] X. He and X. Niyogi, Locality preserving projections, *Advances in Neural Information Processing Systems*, vol.16, pp.153-160, 2004.

[16] X. He, D. Cai, S. Yan et al., Neighborhood preserving embedding, *The 10th IEEE International Conference on Computer Vision*, vol.2, pp.1208-1213, 2005.

[17] D. Cai, X. He and J. Han, Spectral regression: A unified approach for sparse subspace learning, *The 7th IEEE International Conference on Data Mining*, pp.73-82, 2007.

[18] L. Qiao, S. Chen and X. Tan, Sparsity preserving projections with applications to face recognition, *Pattern Recognition*, vol.43, no.1, pp.331-341, 2010.

[19] J. Wang, J. Yang, K. Yu et al., Locality-constrained linear coding for image classification, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3360-3367, 2010.

[20] H. Li, T. Jiang and K. Zhang, Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Networks*, vol.17, no.1, pp.157-165, 2006.

[21] Z. Li, J. Liu, Y. Yang et al., Clustering-guided sparse structural learning for unsupervised feature selection, *IEEE Trans. Knowledge and Data Engineering*, vol.26, no.9, pp.2138-2150, 2014.

[22] A. Rodriguez and A. Laio, Clustering by fast search and find of density peaks, *Science*, vol.344, no.6191, pp.1492-1496, 2014.

[23] J. Demmar, Statistical comparisons of classifiers over multiple data sets, *The Journal of Machine Learning Research*, vol.7, pp.1-30, 2006.