Abstract—The sparse representation-based classifier (SRC) has been developed and shows great potential for pattern classification. This paper aims to gain a discriminative projection such that SRC achieves the optimum performance in the projected pattern space. We use the decision rule of SRC to steer the design of a dimensionality reduction method, which is coined the sparse representation classifier steered discriminative projection (SRC-DP). SRC-DP matches SRC optimally in theory. Experiments are done on the AR and extended Yale B face image databases, and results show the proposed method is more effective than other dimensionality reduction methods with respect to the sparse representation-based classifier.

Keywords—feature extraction, sparse representation, classifier

I. INTRODUCTION

Sparse representation has aroused broad interest in pattern recognition and computer vision areas in the last few years [1]. Recently, Wright et al. presented a sparse representation based classification (SRC) method [2]. The basic idea of SRC is to represent a given test sample as a sparse linear combination of all training samples; the sparse nonzero representation coefficients are supposed to concentrate on the training samples with the same class label as the test sample. The sparse representation-based classification method turns out to be very effective, especially for face recognition.

Sparse representation involves an underdetermined system of linear equation $y = Aw$, where $A$ is an $N \times M$ matrix and $N < M$. For SRC, the columns of matrix $A$ are all training sample vectors. This means that to obtain a sparse solution, the dimension of feature vectors must be smaller than the number of training samples. To deal with small sample problems like face recognition, where the dimension of images is larger than the training sample size, a dimensionality reduction (feature extraction) step becomes necessary before implementing SRC. How should we perform dimensionality reduction such that SRC achieve the optimum performance in the reduced-dimensional space? This problem will be addressed in the paper.

Observing that SRC uses a reconstruction residual based decision rule, we use it to steer the design of a dimensionality reduction method, which is coined the sparse representation classifier steered discriminative projection (SRC-DP). The basic idea of SRC-DP is to seek a linear transformation such that in the transformed low-dimensional space, the within-class reconstruction residual is as small as possible and simultaneously the between-class reconstruction residual is as large as possible. Therefore, SRC can achieve the optimum classification performance in the SRC-DP transformed space.

II. SPARSE REPRESENTATION BASED CLASSIFIER

Suppose there are $c$ known pattern classes. Let $A_i$ be the matrix formed by the training samples of Class $i$, i.e., $A_i = [y_{i1}, y_{i2}, \cdots, y_{iM}] \in R^{d \times M}$, where $M_i$ is the number of training samples of Class $i$. Let us define a matrix $A = [A_1, A_2, \cdots, A_c] \in R^{d \times M}$, where $M$ is the total number of training samples. The matrix $A$ is obviously composed of entire training samples.

Given a test sample $y$, we represent $y$ in a overcomplete dictionary whose basis vectors are training sample themselves, i.e. $y = Aw$. This system of linear equation is underdetermined if $N < M$. The idea of sparse representation based classification is motivated by the following observation: A valid test sample $y$ can be sufficiently represented using only the training samples from the same class. The representation is naturally sparse if train sample size is large enough. The sparser the recovered representation coefficient vector $w$ is, the easier it will be to accurately determine the identity of the test sample $y$ [2].

The sparsest solution to $y = Aw$ can be sought by solving the following optimization problem

$$L_0: \hat{w}_0 = \text{arg } \min ||w||_0, \text{ subject to } Aw = y,$$

where $|| \cdot ||_0$ denotes the $L_0$-norm, which counts the number of nonzero entries in a vector.

Solving $L_0$ optimization problem in Eq.(1), however, is NP hard and extremely time-consuming. Recent research efforts reveal that if the solution $\hat{w}_0$ is sparse enough, finding the solution of the $L_0$ optimization problem is equivalent to finding the solution to the following $L_1$ optimization problem [3, 4]:

$$L_1: \hat{w}_1 = \text{arg } \min ||w||_1, \text{ subject to } Aw = y. $$
This problem can be solved in polynomial time by standard linear programming algorithms \[5\].

After obtaining the sparsest solution \(\hat{w}_1\), we can design a sparse representation based classifier (SRC) in the following way. For each class \(i\), let \(\delta_i(w)\) be a vector whose only nonzero entries are the entries in \(w\) that are associated with class \(i\). Using only the coefficients associated with the \(i\)-th class, one can reconstruct a given test sample \(y\) as \(v^i = A\delta_i(\hat{w}_1)\). \(v^i\) is called the prototype of class \(i\) with respect to the sample \(y\). The distance (residual) between \(y\) and its prototype \(v^i\) of class \(i\) is defined by

\[
r^i(y) = \|y - v^i\|_2 = \|y - A\delta_i(\hat{w}_1)\|_2
\]  

The SRC decision rule is: if \(r^i(y) = \min_{i \neq j} r_j(y)\), \(y\) is assigned to Class \(i\).

### III. SPARSE REPRESENTATION BASED CLASSIFIER

**STEERED DISCRIMINATIVE PROJECTION**

In this section, we will use the classification rule of SRC to induce the design of a dimensionality reduction method, which is coined the sparse representation classifier steered discriminative projection (SRC-DP). The basic idea of SRC-DP is to seek a linear transform \(y = P^T x\) such that the SRC achieves the optimum performance in the reduced-dimensional transformed space.

Let \(B = [B_1, B_2, \cdots, B_c] \in \mathbb{R}^{N \times M}\) be the training data matrix in the original input space, where \(B_i = [x_{i1}, x_{i2}, \cdots, x_{im}] \in \mathbb{R}^{N \times mi}\) is the matrix formed by the training samples of Class \(i\). Under a given linear transform \(y = P^T x\), each data point \(x_j\) in input space \(\mathbb{R}^{N}\) is mapped into \(y_j = P^T x_j\) in a \(d\)-dimensional space \(\mathbb{R}^d\). As a result, the data matrix in the original input space is converted into the one in \(\mathbb{R}^d\), that is, \(A = P^T B\).

Let us now consider the classification problem using SRC in the mapped \(d\)-dimensional space. For each training sample \(y_j\), let us leave it out from the training set and use the remaining training samples to linearly represent it. By solving the \(L_1\) optimization problem in Eq. (2), we obtain a representation coefficient vector \(w_j\). Let \(\delta_j(w_j)\) be the representation coefficient vector with respect to Class \(s\). Then, the prototype of class \(s\) with respect to the sample \(y_j\) is \(v^s_j = A\delta_j(w_j)\), \(s = 1, \cdots, c\). The distance between \(y_j\) and Class \(s\) is defined as

\[
d_s(y_j) = \|y_j - v^s_j\|^2
\]  

Considering the decision rule of SRC, to make the classifier perform well, we expect that \(y_j\) is as close as possible to its within-class prototype \(v^s_j\) and simultaneously as far away as possible from its between-class prototypes \(v_{ij}\) \((s \neq i)\). That is, the within-class distance \(d_s(y_j)\) is supposed to be as small as possible and the between-class distance \(d_i(y_j)\) for each \(s \neq i\) as large as possible. To make SRC achieve good performance on all training samples, we would like to characterize the average within-class and between-class distances, i.e., the within-class and between-class scatters. Specifically, the within-class scatter is defined as follows

\[
\overline{S}_w^L = \frac{1}{M} \sum_{i,j} d_i(y_j) = \frac{1}{M} \sum_{ij} \|y_j - v^s_j\|^2 = \text{tr}(\overline{S}_w^L),
\]  

where \(\overline{S}_w^L = \frac{1}{M} \sum_{i,j} (y_j - v^s_j)(y_j - v^s_j)^T\) is the within-class scatter matrix in the transformed space and \(\text{tr}(\cdot)\) is the trace operator.

The between-class local scatter of samples is defined as follows

\[
\overline{S}_b^L = \frac{1}{M(c-1)} \sum_{i,j} d_i(y_j) = \frac{1}{M(c-1)} \sum_{ij} \|y_j - v^s_j\|^2 = \text{tr}(\overline{S}_b^L),
\]  

where \(\overline{S}_b^L = \frac{1}{M(c-1)} \sum_{i,j} (y_j - v^s_j)(y_j - v^s_j)^T\) is the between-class scatter matrix in the transformed space.

According to the SRC decision rule, larger between-class scatter and smaller within-class scatter will lead to better classification results in an average sense. Therefore, we can choose to maximize the following criterion function

\[
J(P) = \frac{\text{tr}(\overline{S}_b^L)}{\text{tr}(\overline{S}_w^L)}
\]  

Now, let us consider how to determine a projection matrix \(P\) based on the criterion. Towards this end, we try to convert the criterion to be a function with respect to \(P\).

Inserting \(y_j = P^T x_j\), \(A = P^T B\) and \(v^s_j = A\delta_j(w_j)\) into the formula of \(\overline{S}_w^L\) and \(\overline{S}_b^L\), respectively, we have

\[
\overline{S}_w = P^T S_w^L P,
\]  

\[
\overline{S}_b = P^T S_b^L P,
\]  

\[
J(P) = \frac{\text{tr}(P^T S_b^L P)}{\text{tr}(P^T S_w^L P)}
\]  

695
\[
\tilde{S}_k^L = P^T S_k^L P, \tag{9}
\]

where the two matrices \(S_k^L\) and \(S_k^b\) are defined as

\[
S_k^L = \frac{1}{M} \sum_{ij} [x_{ij} - B \delta_i (w_{ij})] ([x_{ij} - B \delta_i (w_{ij})]^T, \tag{10}
\]

\[
S_k^b = \frac{1}{M (e-1)} \sum_{ij} \sum_{j \neq i} [x_{ij} - B \delta_i (w_{ij})] ([x_{ij} - B \delta_i (w_{ij})]^T. \tag{11}
\]

\(S_k^L\) and \(S_k^b\) are called the within-class and between-class sparse scatter matrices, respectively.

The criterion in Eq. (7) thereby becomes

\[
J(P) = \frac{\text{tr}(P^T S_k^L P)}{\text{tr}(P^T S_k^b P)}. \tag{12}
\]

If the two matrices \(S_k^L\) and \(S_k^b\) can be constructed directly in the original input space, the optimal projection matrix \(P\) can be determined by maximizing the criterion in Eq. (12). Actually, the column vectors of the optimal projection matrix can be chosen as the generalized eigenvectors of \(S_k^L \phi = \lambda S_k^b \phi\) associated with \(d\) largest eigenvalues. However, the representation coefficient vector \(w_{ij}\) in the formula of \(S_k^L\) and \(S_k^b\) is unknown to us without the projection matrix \(P\) being given in advance, noticing that \(w_{ij}\) is calculated in the transformed space \(R^d\).

In summary, given an initial projection matrix \(P_0\) (which can be chosen as a random matrix) in advance, we can map the data \(x_{ij}\) in the original space into \(y_{ij} = P_0^T x_{ij}\) in the transformed space. Representing each point \(y_{ij}\) using the remaining training samples and obtain the representation coefficient vector \(w_{ij}\) by solving the \(L_1\) optimization problem. We then can construct the matrices \(S_k^L\) and \(S_k^b\) in the original space and obtain a new projection matrix \(P_k\) by solving the corresponding generalized eigenvalue problem. Based on this idea, we can derive an iterative SRC-DP algorithm. In the \(k\)-th iteration of the algorithm, we use \(P_{k-1}\), the resulting projection matrix after the \((k-1)\)-th iteration, as the initial projection matrix to input the system to yield a new projection matrix \(P_k\). The iteration process continues until the algorithm converges. Convergence may be determined by observing when the value of the criterion function \(J(P)\) in Eq. (20) stops changing. Specifically, after \(k\) times of iterations, if \(|J(P_k) - J(P_{k-1})| < \epsilon\), we think the algorithm converges and choose \(P^* = P_k\). To alleviate the effect of the magnitude of the value of \(J(P)\) and make it easy to determine a proper \(\epsilon\), we use the relative difference based convergence criterion instead here, that is, \(|J(P_k) - J(P_{k+1})|/|J(P_k)| < \epsilon\).

The SRC-DP algorithm is illustrated in Fig. 1.

IV. EXPERIMENTS

A. Experiment Using the AR Database

The images of 120 individuals (65 men and 55 women) of the AR database were taken in two sessions (separated by two weeks) and each section contains 13 color images. 14 face images without occlusions (each session containing 7) of these 120 individuals are selected and used in our experiment. The face portion of each image is manually cropped and normalized to 50×45 pixels. The normalized images of one person are shown in Fig. 2, where (a)–(g) are from Session 1, and (n)–(t) are from Session 2. The details of the images are: (a) neutral expression, (b) smile, (c) anger, (d) scream, (e) left light on; (f) right light on; (g) all sides light on; and (n)–(t) were taken under the same conditions as (a)–(g).

In our experiment, images from the first session were used for training, and images from the second session were used for testing. PCA (Eigenfaces [7]), FLDA (Fisherfaces [7]), LPP (Laplacianfaces [8]) and the proposed SRC-DP method are, respectively, used for feature extraction. To avoid overfitting, we first perform PCA and reduce the dimension to be 200 before implementing FLDA, LPP and SRC-DP. Finally, the SRC classifier is employed for classification. The maximal recognition rate of each method and the corresponding dimension is listed in Table I. Table I
shows that the maximal recognition rate of SRC-DP is significantly better than those of the other methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCA</th>
<th>FLDA</th>
<th>LPP</th>
<th>SRC-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate</td>
<td>69.0</td>
<td>78.9</td>
<td>70.0</td>
<td>83.3</td>
</tr>
<tr>
<td>Dimension</td>
<td>110</td>
<td>190</td>
<td>150</td>
<td>150</td>
</tr>
</tbody>
</table>

B. Experiment using the Extended Yale B database

The extended Yale B face database [9] contains 38 human subjects under 9 poses and 64 illumination conditions. All frontal-face images marked with P00 are used in our experiment. Each image is resized to 42×48 pixels and further pre-processed by histogram equalization.

<table>
<thead>
<tr>
<th>Training sample size</th>
<th>PCA</th>
<th>FLDA</th>
<th>LPP</th>
<th>SRC-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>65.3</td>
<td>63.9</td>
<td>52.9</td>
<td>72.1</td>
</tr>
<tr>
<td>16</td>
<td>79.6</td>
<td>80.9</td>
<td>77.4</td>
<td>91.2</td>
</tr>
<tr>
<td>24</td>
<td>94.0</td>
<td>98.5</td>
<td>95.1</td>
<td>98.9</td>
</tr>
<tr>
<td>32</td>
<td>98.8</td>
<td>99.0</td>
<td>98.4</td>
<td>99.0</td>
</tr>
</tbody>
</table>

In our test, we use the first 8, 16, 24, 32 images per subject, respectively, for training, and the remaining images for testing. PCA, FLDA, LPP and the proposed SRC-DP method are used for feature extraction. Before implementing FLDA, LPP and SRC-DP, we use PCA to reduce the dimension to be 100, 120, 160 and 200 with respect to different number of training samples per class. Finally, the SRC classifier is employed for classification. The maximal recognition rate of each method is listed in Table II. Table II shows that when the training sample size is relative small, SRC-DP significantly outperforms PCA, FLDA and LPP. When the training sample size becomes large, SRC-DP still achieves comparable results with the other methods.

V. CONCLUSIONS

This paper addresses the problem on how to perform dimensionality reduction such that the sparse representation-based classifier (SRC) achieves the optimum performance in the reduced dimensional space. We use the decision rule of SRC to steer the design of a new dimensionality reduction method which matches SRC optimally in theory. Experiments are done on the AR and Yale B face image databases, and results demonstrate the proposed method is more effective than PCA, LDA and LPP with respect to the sparse representation-based classifier.

ACKNOWLEDGMENT

This work was partially supported by the Program for New Century Excellent Talents in University of China, the NUST Outstanding Scholar Supporting Program, the National Science Foundation of China under Grants No. 60973098, 60632050 and 90820306, and the NUS Research Grant R-146-000-087-112.

REFERENCES


Figure 2 Samples of the cropped images of one person in the AR database