Short Papers

Face Recognition via Collaborative Representation: Its Discriminant Nature and Superposed Representation

Weihong Deng, Jiani Hu, and Jun Guo

Abstract—Collaborative representation methods, such as sparse subspace clustering (SSC) and sparse representation-based classification (SRC), have achieved great success in face clustering and classification by directly utilizing the training images as the dictionary bases. In this paper, we reveal that the superior performance of collaborative representation relies heavily on the sufficiently large class separability of the controlled face datasets such as Extended Yale B. On the uncontrolled or undersampled dataset, however, collaborative representation suffers from the misleading coefficients of the incorrect classes. To address this limitation, inspired by the success of linear discriminant analysis (LDA), we develop a superposed linear representation classifier (SLRC) to cast the recognition problem by representing the test image in term of a superposition of the class centroids and the shared intra-class differences. In spite of its simplicity and approximation, the SLRC largely improves the generalization ability of collaborative representation, and competes well with more sophisticated dictionary learning techniques, on the experiments of AR and FRGC databases. Enforced with the sparsity constraint, SLRC achieves the state-of-the-art performance on FERET database using single sample per person.

Index Terms—Sparse representation, collaborative representation, sparse subspace clustering, face clustering, face recognition

1 Introduction

A fundamental assumption on image representation is that an image can be encoded in term of a linear superposition of an ensemble of basis images. The image code is determined by the choice of basis images. The goal of efficient coding is to find a set of basis images, which spans the image space, and results in the coefficient values being as uncorrelated or independent as possible over an ensemble of training images [1]. One line of approach to this problem is based on principal component analysis, well-known as Eigenfaces in computer vision community, which aims to find a set of mutually orthogonal basis images that capture the direction of maximum variance in the face space and for which the coefficients are pairwise uncorrelated [2]. Eigenfaces is an unsupervised coding method for reconstruction but not for discrimination [3].

Started by the influential SRC [4], collaborative representation (CR) based approaches have achieved surprisingly good performance on face clustering [5] and classification [4]. They directly utilize the training images themselves as the basis images, and assume that the unseen sample can be linearly represented by the training samples in the same class. Based on the coding coefficients spanned by all training samples from all classes, CR based methods expect that the major components can be found in the correct class. Although previous studies have validated that the coefficient regularizer is not crucial [6], [7], it is still unclear why the (unsupervised) coding coefficients based classification, such as SRC and CRC, can outperform the state-of-the-art (supervised) classifier such as SVM in face recognition.

In this paper, we reveal that the discriminant nature of the collaborative representation is determined by the class separability of the data dictionary, measured by the quantity $J = \mathrm{Tr}\{S_T^{\dagger} S_B\}$ involving the inter-class scatter normalized by the global scatter. On the controlled face datasets, because that the class separability is sufficiently large for nearly perfect clustering and classification, and the coding coefficients of the CR become naturally discriminative. This class separability of the facial images can be approximately exploited by data whitening or discriminant analysis. As evidence, we evaluate two *baseline algorithms* based on the simple metrics that characterize the quantity of $\mathrm{Tr}\{S_T^{\dagger} S_B\}$, followed by the traditional clustering or classification methods. Empirical results show that both the CR based methods, such as SSC [5] and SRC [4], and our proposed baseline methods can take advantage of the large class separability of controlled dataset to obtain excellent clustering and classification performance.

Unfortunately, on the uncontrolled and undersampled datasets, CR based methods suffer from the misleading coding coefficients of the incorrect classes. To address this limitation, we propose to decompose the training sample of CR into prototype (class centroid) and variation (sample-to-centroid difference) parts, and propose a superposed linear representation that encodes the test sample as a superposition of the prototype and variation bases [8]. Experimental results on AR, FRGC, and FERET databases show that the proposed SLRC achieves better performance than current sophisticated dictionary learning methods, using the undersampled and uncontrolled training data. Furthermore, enforced with the sparsity constraint, SLRC achieves state-of-the-art single-sample based face recognition performance using an overcomplete variation dictionary.

2 Discriminant Nature of Collaborative Representation

This section introduces our finding that the collaborative representation is discriminative because the controlled dataset has sufficiently large class separability, and this class separability can be equivalently exploited by traditional feature extraction techniques such as data whitening and discriminant analysis.

2.1 Problem Definition

Given the training samples denoted by a matrix $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$ and a test sample denoted by a vector $y \in \mathbb{R}^{d \times 1}$, we consider the basic problem of representing the test image y as a linear combination of the training image ensemble, i.e., $y = X\alpha$. By assuming that training samples have been projected into low-dimensional feature spaces, the coefficient vector α is underspecified, i.e., many choices of α lead to the same y. To avoid the complex effect induced by regularization, we analyze the characteristics of coefficients of the least-norm solution. Specifically, the least-square solution considers the optimization problem

$$\min \|\alpha\|_2, \text{ s.t. } y = X\alpha, \tag{1}$$

where optimal solution $\alpha = X^T(XX^T)^{-1}y$ has the smallest norm of any solution.

In mathematics, although the condition of feature matrix X varies, an unique generalized solution to the CR model always exists such that the squared reconstruction error $\|y - X\alpha\|_2$ and the

• *The authors are with School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: {whdeng, jnhu, guojun}@bupt.edu.cn.*

Manuscript received 28 Sept. 2016; revised 15 July 2017; accepted 20 Sept. 2017. Date of publication 28 Sept. 2017; date of current version 12 Sept. 2018.
(Corresponding author: Weihong Deng.)
Recommended for acceptance by M. Tistarelli.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2757923

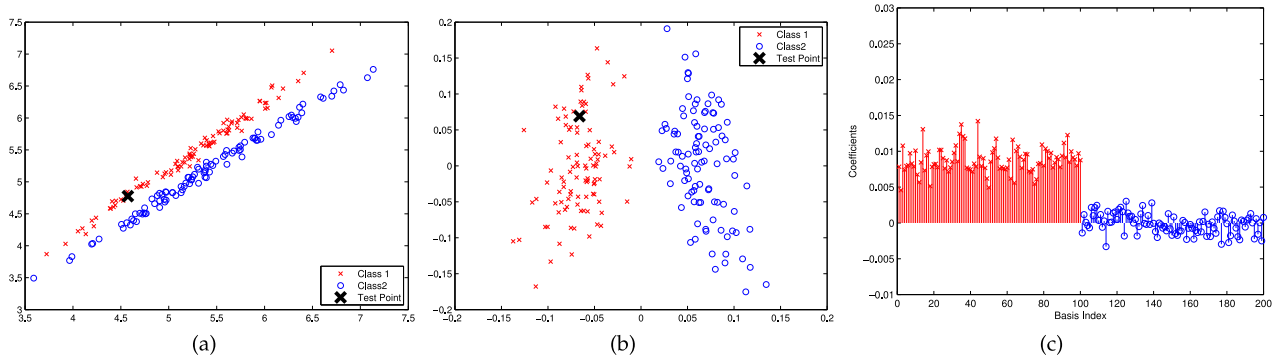


Fig. 1. Two-class examples with identical class separability, i.e., $\mathrm{Tr}\{S_T^{\dagger}S_B\}$, where the points marked by crosses are the test samples to be reconstructed. (a) The example with subtle inter-class variance, relative to the large total variance. Fortunately, the inter-class variance is perpendicular to the principal direction of global variance. (b) The example with large inter-class variance, relative to the total variance. (c) The coding coefficients of two test samples in (a) and (b) are identical, and very discriminative.

squared norm of the solution $\alpha^t\alpha$ are both minimized.[1] This unified and unique solution is denoted by

$$\alpha = X^{\dagger}y, \tag{2}$$

where $X^{\dagger}$ is called the Pseudoinverse [9] of X. In particular, when X is full column rank, $X^{\dagger}$ is computed as $X^{\dagger} = (X^TX)^{-1}X^T$ with $X^{\dagger}X = I$. When X has full row rank, $X^{\dagger}$ is computed as $X^{\dagger} = (X^T(XX^T)^{-1})$ with $XX^{\dagger} = I$.

2.2 Close Relationship to Class Separability

The discriminant power of CR comes from the coding coefficients, and we analyze their characteristic by evaluating the class-specific summation of the coefficients. Specifically, we sum up the coding coefficients associated with the ith class, $s_i = \sum_{j=1}^{n_i}\alpha_{i,j}$, and evaluate their deviation from the ideal class-specific summation. For the simplicity of analysis, the ideal class-specific summation of class i is set to 1 for the samples of the ith class, and set to 0 otherwise. We define the discriminant power of the coefficients as the consistency with this ideal case.

Formally, the discriminatory ability of the coding coefficients of sample x can be measured by squared error between ideal and real concentration as follows:

$$e = \|I_i - TX^{\dagger}x\|^2, \text{ if } x \in \omega_i, \tag{3}$$

while $I_i = [0,\ldots,1,\ldots,0]^t \in \mathbb{R}^C$ is a class indictor vector whose only nonzero entry is the ith entry. For the whole training set X, one can construct an indicator matrix $T \in \mathbb{R}^{C\times n}$ whose nonzero entry in each column indicates the class label of each sample. Finally, the discriminatory ability of the collaborative coefficients of the whole training set could be measured by the sum of "deviation from ideal concentration" over all samples as follows:

$$E = \|T - TX^{\dagger}X\|^2, \tag{4}$$

To simplify the analysis, we assume the training data are centered, and denoted as $\hat{X}$, i.e., $\hat{X}^T1 = 0$, and concentration degree of the coefficients can be analyzed as follows:

$$\begin{aligned} E &= \|T - T\hat{X}^{\dagger}\hat{X}\|^2 \\ &= \mathrm{Tr}\{(\hat{T} - \hat{T}\hat{X}^{\dagger}\hat{X})(\hat{T} - \hat{T}\hat{X}^{\dagger}\hat{X})^t\} \\ &= \mathrm{Tr}\{\hat{T}\hat{T}^t - \hat{T}\hat{X}^{\dagger}\hat{X}\hat{T}^t - \hat{T}\hat{X}^t(\hat{X}^{\dagger})^t\hat{T}^t + \hat{T}\hat{X}^{\dagger}\hat{X}\hat{X}^t(\hat{X}^{\dagger})^t\hat{T}^t\} \\ &= \mathrm{Tr}\{\hat{T}\hat{T}^t - \hat{T}\hat{X}^t(\hat{X}\hat{X}^t)^{\dagger}\hat{X}\hat{T}^t\} \\ &= \mathrm{Tr}\{\hat{T}\hat{T}^t\} - \mathrm{Tr}\{(\hat{X}\hat{X}^t)^{\dagger}\hat{X}\hat{T}^t\hat{T}\hat{X}^t\} \\ &= \mathrm{Tr}\{\hat{T}\hat{T}^t\} - \mathrm{Tr}\{S_T^{\dagger}S_B\}. \end{aligned} \tag{5}$$

Note that the matrix $\hat{X}\hat{X}^t$ is the total scatter matrix S_T, and $\hat{X}\hat{T}^t\hat{T}\hat{X}^t$ is the between-class scatter matrix, S_B. Thus, since the target matrix T is fixed, the minimum value of E is determined by the trace of $\mathrm{Tr}\{S_T^{\dagger}S_B\}$, which is widely used in discriminant analysis to measure the class separability of the data. It is an intrinsic property of data themselves, regardless of the solvers of the least-square problem.

When the CR model is applied on the data set with large $\mathrm{Tr}\{S_T^{\dagger}S_B\}$, the least-square coding coefficients are naturally concentrate on the correct class, and thus the algorithms based on these coefficients, such as SRC and SSC, also naturally become discriminative, without any special regularization. Our recent work obtains excellent recognition accuracy by simply accumulating the coefficients of each class [10]. Fig. 1 shows two examples, i.e., (a) and (b), where both data sets have the same large $\mathrm{Tr}\{S_T^{\dagger}S_B\}$, and the two samples marked by the dark cross have identical coding coefficients shown in (c). One can see from the Fig 1c that the coding coefficients are *dense but discriminative*: all the large coefficients concentrate on the correct class. Indeed, regularized model may generate more sparse or concentrated coefficients, but the resulting complex computation might be not necessary.

2.3 Geometric Interpretation

It is easy to understand that least-square coding coefficients are discriminative where the data classes are distributed far apart as in Fig. 1b. For face processing, however, it is a common knowledge that "the variations between the images of the same class due to illumination are almost always larger than image variations due to change in class"[11]. Why is the CR model still applicable for both face clustering and classification? To investigate this question, we analyze the physical meaning of the quantity $\mathrm{Tr}\{S_T^{\dagger}S_B\}$ by its spectral decomposition. Specifically, let $U_B = \{u_{b1},\ldots,u_{bq}\}$ and $\Lambda_B = \{\lambda_{b1},\ldots,\lambda_{bq}\}$ be the eigenvector and eigenvalues of $S_BU_B = U_B\Lambda_B$, and $U_T = \{u_{t1},\ldots,u_{tp}\}$ and $\Lambda_T = \{\lambda_{t1},\ldots,\lambda_{tp}\}$ be the eigenvector and eigenvalues of $S_TU_T = U_T\Lambda_T$, where q and p are the ranks of S_B and S_T respectively, $\lambda_{b1} \geq \lambda_{b2} \geq \ldots \geq \lambda_{bq}$, $\lambda_{t1} \geq \lambda_{t2} \geq \ldots \geq \lambda_{tp}$, and $p \geq q$. In light of the similar formulation in [12], the quantity can be decomposed as follows:

$$\mathrm{Tr}\{S_T^{\dagger}S_B\} = \sum_{i=1}^{q}\sum_{j=1}^{p}\frac{\lambda_{bi}}{\lambda_{tj}}\left(u_{tj}^Tu_{bi}\right). \tag{6}$$

The spectral decomposition of the scatter matrices reveals that the discriminant power of the coding coefficients is determined by the sum of the inter-class variances normalized (divided) by the total variances along the consistent directions. In other words, on the cases where the inter-class variance is small but the total variance is also small along the that direction, CR model can recover the cluster separability hidden in the high dimensional space, such as that in Fig. 1a.

In a typical controlled face database, the inter-class variance comes from the subtle difference of the local texture and shape around the facial features, but the intra-class variance is mainly

1. The side conditions used to define the Moore-Penrose pseudo-inverse are that the squared representation error be minimized and, if there is ambiguity (several solutions with the same minimum error), the ℓ_2 norm of α also be minimized.

caused by global appearance change of illumination and expression. In the image space, the intra-class image differences are approximately uncorrelated to the inter-class image differences [13]. This property of the inter-class and intra-class scatters makes the principal basis of S_B and those of S_T are not conflict in the high dimensional image space [12], so that the quantity of $\mathrm{Tr}\{S_T^{\dagger}S_B\}$ is sufficiently large. In this sense, *the subtle inter-class variance can be highlighted in the background of dominant intra-class variance*. In our experimental study, we will explore this desirable characteristic for the face clustering and classification using some $\mathrm{Tr}\{S_T^{\dagger}S_B\}$ related metrics, and further compare them with CR based methods.

3 Superposed Linear Representation based Classification (SLRC)

Inspired by the decomposed representation in discriminant analysis, this section introduces a superposed linear representation model that constructs dual dictionaries to separately exploit the inter-class and intra-class variability for CR based classification.

3.1 Decomposed Representation of Linear Discriminant Analysis

Linear discriminant techniques that aim to preserve class separability have achieved great success in face recognition [14], [15]. Further, by handling the inter- and intra-class variations separately, previous studies have reported very successful face-recognition results using the Bayesian matching [13] and unified subspace analysis [16] framework. Given a data set with multiple samples per class, the n_i samples of class i form a matrix $X_i \in \mathbb{R}^{d\times n_i}$, $i = 1,\ldots,k$, $\sum_{i=1}^{k} n_i = n$. Considering the class labels, we introduce three $d \times n$ basis matrices

$$H_W = [X_1 - c_1 e_1^T, X_2 - c_2 e_2^T, \ldots, X_k - c_k e_k^T], \tag{7}$$

$$H_B = [(c_1 - c)e_1^T, (c_2 - c)e_2^T, \ldots, (c_k - c)e_k^T], \tag{8}$$

$$H_T = \hat{X} = [(x_1 - c), \ldots, (x_n - c)] = X - ce^T, \tag{9}$$

where $e_i = [1,\ldots,1]^T \in \mathbb{R}^{n_i\times 1}$, $e = [1,\ldots,1]^T \in \mathbb{R}^{n\times 1}$, $c_i = \frac{1}{n_i}X_i e_i$ is the geometric centroid of class i, and $c = \frac{1}{n}X_i e$ is the global centroid. Interestingly, the basis matrices have the relationship $H_T = H_W + H_B$ [17].

The classical PCA technique derives the subspace by the eigenvectors of the total scatter matrix $S_T = H_T H_T^T$, which is optimal for information-preserving and helpful for removing the unreliable dimension [18]. In contrast, LDA tries to seek the subspace that best discriminates different classes by maximizing the between-class scatter, while minimizing the within-class scatter in the projective subspace. In the theory of LDA, between-class scatter matrix $S_B = H_B H_B^T$ characterizes the relation between any two class centroids. The within-class scatter matrix $S_W = H_W H_W^T$ characterizes sample variations deviation from corresponding class centroid. In this respect, LDA essentially first decomposes the centered data into two parts as $\hat{X} = H_B + H_W$, and then find the projective bases by the optimization criterion

$$J(w) = \max \frac{w^T H_B H_B^T w}{w^T H_W H_W^T w}. \tag{10}$$

In the sense, we denote the relationship $\hat{X} = H_B + H_W$ as the *decomposed representation* of LDA: H_B is an approximated representation that characterizes the samples by corresponding class centroids, and H_W represents the residuals of each sample deviated from the approximation H_B.

3.2 Superposed Linear Representation Based Classification

Although having achieved the great success in robust face recognition [4], [7], CR suffers from the undersampled problem: When the training images are insufficient or unrepresentative, the test sample has to be reconstructed by the samples of other classes, and thus the coding coefficients generate misleading results. In essence, this problem is caused by the mixture of the inter-class and intra-class components in the dictionary bases, where the intra-class components of the testing image are possibly borrowed from the incorrect identities. To overcome this difficulty, we attempt to decompose the collaborative dictionary in a manner similar to the decomposed representation in LDA inspired by its success in undersampled classification. Specifically, given a sample x from one of the classes in the training set, we assume it can be naturally reconstructed by two parts

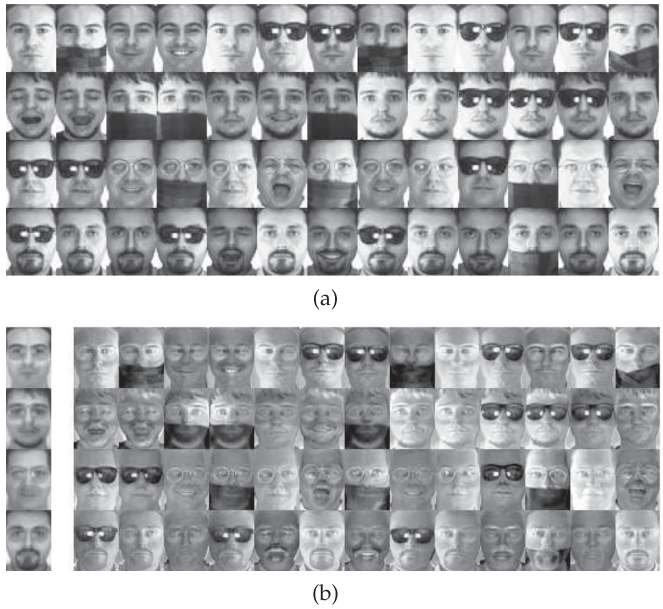
$$x = c_{(x)} + (x - c_{(x)}), \tag{11}$$

where $c_{(x)}$ is the centroid of corresponding class, and $x - c_{(x)}$ is the intra-class difference from the sample to its class centroid. Applying this "naive" decomposition to each training sample, we decompose the dictionary of CR into prototype and variation dictionaries. Following previous notations, the prototype dictionary can be represented as follows:

$$P = [c_1, \ldots, c_i, \ldots, c_k] \in \mathbb{R}^{d\times k}, \tag{12}$$

where c_i is the centroid of class i. As the prototypes are represented by class centroids, the variation dictionary is naturally constructed by the sample based difference to the centroids as follows:

$$V = H_W = [X_1 - c_1 e_1^T, \ldots, X_k - c_k e_k^T] \in \mathbb{R}^{d\times n}. \tag{13}$$



(a)

(b)

Fig. 2. The illustrative examples of the "prototype plus variation" superposed linear representation model. (a) the randomly selected training images from AR database. (b) the first column contains the "prototypes" derived by averaging the images of the same subject, and the rest columns are the "sample-to-centroid" variation images.

Fig. 2 illustrates an typical example of the prototype and variation dictionaries in the image form. One can see from the figure that the class centroids are visualized as stabilized average images [19], [20], and the variation images separate out the uncontrolled factors, such as lighting and sunglasses. With the prototype and variation dictionaries, we propose the Superposed Linear Representation-based Classification that casts the recognition problem as finding a linear representation of the test image in term of a superposition of the class centroids and the intra-class differences. It is interesting to point out the similarities between LDA and SLRC as follows.

- Both the prototype dictionary and H_B of LDA use an approximated representation that characterizes the samples by corresponding class centroids.

- The variation dictionary is identical to H_W of LDA, which is designed to represent the residuals of each samples deviated from the centroid based approximation. Both $S_W = H_W H_W^T$ of LDA and the variation dictionary are shared across all classes.

While LDA emphasizes mainly on discriminative dimension reduction, SLRC simultaneously satisfies the needs for adequate signal reconstruction and subsequent classification performance using dual decomposed dictionaries, which provides a flexible CR by imposing various regularization on the coefficients. Algorithm 1 below summarizes the complete recognition procedure.

Algorithm 1. Superposed Linear Representation based Classification(SLRC)

1: **Input**: a matrix of training samples $A = [A_1, A_2, \ldots, A_k] \in \mathbb{R}^{d \times n}$ for k classes, and an regularization parameter $\lambda > 0$. Compute the prototype matrix P according to (12), and the variation matrix V according to (13). When the sample size per class is insufficient, the matrix V can be supplemented from a set of generic samples outside the gallery.

2: Compute the projection matrix $\Phi \in \mathbb{R}^{d \times p}$ by applying PCA on the training samples A, and project the prototype and variation matrices to the p-dimensional space.

$$P \leftarrow \Phi^T P,\ V \leftarrow \Phi^T V. \tag{14}$$

3: Normalize the columns of P and V to have unit ℓ_2-norm, and solve the ℓ_1 or ℓ_2-minimization problem

$$\begin{bmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \end{bmatrix} = \arg\min \left\| [P, V] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - y \right\|_2^2 + \lambda \left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_\ell, \tag{15}$$

where $\alpha, \hat{\alpha} \in \mathbb{R}^k$, $\beta, \hat{\beta} \in \mathbb{R}^n$. The norm of coefficients $\ell \in \{1, 2\}$ in our experiment, and corresponding algorithms are denoted as SLRC-ℓ_1 and SLRC-ℓ_2 respectively.

4: Compute the residuals

$$r_i(y) = \left\| y - [P, V] \begin{bmatrix} \delta_i(\hat{\alpha}_1) \\ \hat{\beta}_1 \end{bmatrix} \right\|_2, \tag{16}$$

for $i = 1, \ldots, k$, where $\delta_i(\hat{\alpha}_1) \in \mathbb{R}^n$ is a new vector whose only nonzero entries are the entries in $\hat{\alpha}_1$ that are associated with class i.

5: **Output**: $Identity(y) = \arg\min_i r_i(y)$.

When the number of samples per class is insufficient, and in particular when only a single sample per class is available, the intra-class variation matrix would become collapsed. To address this difficulty, one can acquire the intra-class variation bases from the generic subjects outside the gallery, which are assumed to be shareable across different subjects.

There have been a number of dictionary learning methods [21], [22], [23], [24], [25], [26] that effectively improve the generalization ability of CR. The most similar method is the SDR-SLR [27] that applies class-wise low-rank decomposition to separate the identity and intra-class variation dictionaries, and derives sparse and dense coefficients for two dictionaries respectively. Compared with SDR-SLR and other learning methods, the "naive" centroid-based dictionary decomposition of SLRC is much more simple, efficient, and parameter-free. Actually, SLRC has not induced any new parameter compared to the classical SRC. Although the class centroid is generally an approximated representation, SLRC competes well with more sophisticated dictionary learning techniques in our experiments.

4 Experimental Study

In this section, we first perform the study on the controlled database to analyze the relationship between class separability

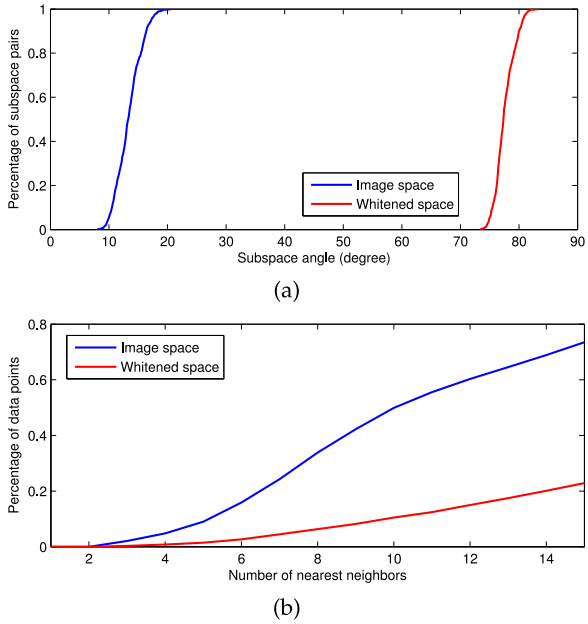


Fig. 3. (a) Percentage of pairs of subspaces whose smallest principal angle is smaller than a given value. (b) Average percentage of data points in pairs of subspaces that have one or more of their K-nearest neighbors in the other subspace.

$\mathrm{Tr}\{S_T^\dagger S_B\}$ and CR methods. Then, we further demonstrate the effectiveness of the proposed SLRC on general recognition experiments of the AR, FRGC, and FERET database with uncontrolled and undersampled training datasets.

4.1 Face Clustering on Controlled Dataset

The Extended Yale B dataset consists of 192×168 pixel cropped face images of 38 individuals, where there are around 64 frontal face images for each subject acquired under controlled lighting conditions. To reduce the computational cost and the memory requirements of all algorithms, we downsample the images to 48×42 pixels and treat each 2,016D vectorized image as a data point. As the images are captured under strictly controlled lightning conditions, it has been validated that the images of each class approximately reside in a 9-dimensional subspace [11]. In light of the quantity of $\mathrm{Tr}\{S_T^\dagger S_B\}$, the class separability can be directly measured by between-class scatter, if the data is whitened, i.e., $S_T = I$. Therefore, to better understand this controlled dataset, we first present some geometric statistics before and after whitening process.

First, we compute the smallest principal angle for each pair of subspaces, and accumulate the percentage of the subspace pairs whose smallest principal angle is below a certain value, ranging from 0 to 90 degrees. Fig. 3a shows that the subspaces before and after whitening process have dramatically different principal angles. Before whitening, principal angles between subspaces are between 10 and 20 degrees, which indicates that the data between different subspaces are highly consistent and correlated in the image space. In contrast, in the whitened space, principal angles between subspaces are always larger than 75 degrees.

Second, for each pair of subspaces, we accumulate the percentage of data points that have one or more of their K-nearest neighbors in the other subspaces. As shown in Fig. 3b, a large proportion of data points have the nearest neighbors come from the other subspace, and this percentage rapidly increases as the number of nearest neighbors increases. This clearly shows that the within-class variance is much larger than the between-class variance, and the clustering of such data is a challenging task. In contrast, in the whitened space, there are much fewer nearest neighbors belong to other subspaces. This observation is consistent with our previous

TABLE 1
Clustering Error (%) of Different Algorithms on the Extended Yale B Database

Algorithm	LSA [31]	SCC [32]	LRSC [33]	LRR-H [34]	SSC [5]	**WSC**
5 Subjects						
Mean	58.02	58.90	12.24	6.90	4.31	**3.75**
Median	56.87	59.38	11.25	5.63	**2.50**	3.44
8 Subjects						
Mean	59.19	66.11	23.72	14.34	5.85	**4.28**
Median	58.59	64.65	28.03	10.06	4.49	**3.52**
10 Subjects						
Mean	60.42	73.02	30.36	22.92	10.94	**4.69**
Median	57.50	75.78	28.75	23.59	5.63	**3.59**

studies [28], [29], [30] that showed the whitening process largely enhanced the class separability for face recognition.

In the following, we perform clustering experiment on this controlled face dataset as detailed in [5]. Our clustering baseline, called whitened spectral clustering (WSC), applies conventional spectral clustering in the whitened PCA subspace. Specifically, WSC first applies whitened PCA by retaining 98 percent data variance, and then builds the 7-NN graph for spectral clustering where the affinity matrix is calculated as follows.

$$S(z_i, z_j) = \exp\left(-\frac{1 - cos(z_i, z_j)}{2 \times (0.33)^2}\right). \tag{17}$$

The parameters of other methods follow the reference [5]. The comparative clustering results are shown in Table 1, and one can see from the table that the WSC baseline obtains lower clustering errors than the state-of-the-art subspace clustering algorithms. The average clustering error rates are as low as 3.75 percent and 4.69 percent average clustering error for 5 and 10 subjects, respectively. This excellent clustering performance of WSC and the geometric findings in Figs. 3a and b are consistent. From the nearly perfect clustering accuracy, one can conjecture that this controlled dataset indeed has large class separability, measured by $\mathrm{Tr}\{S_T^{\dagger} S_B\}$. For this reason, the whitening process is able to dramatically improve the subspace separability, and the conventional spectral clustering method in the whitened space can achieve excellent performance. Although CR based methods, such as SSC and LRR, also take advantage of the large class separability, their coding coefficients are suboptimal to measure the neighborhood closeness as indicated by the higher clustering errors.

4.2 Face Classification on Controlled Dataset

This experiment strictly follows the experiment on the Extended Yale B database in the influential paper [4], which concludes that SRC outperforms the state-of-the-art classifiers, such as linear support vector machine (L-SVM) [35] and nearest subspace (NS) [36], on various feature spaces. As in [4], we randomly select 32 images for training for each subject (i.e., about a half of the images per subject) and the other images for testing. Three conventional features, namely Eigenfaces, Laplacianfaces, Fisherfaces, and two unconventional features, namely downsampled images and Randomfaces, are tested. Following the experiment in [4], we compute the recognition rates with feature space dimensions 30, 56, 120, 504. Note that Fisherfaces is only available at dimension 30 limited by the number of classes. To preserve the class separability, our baseline algorithm, called linear discriminant analysis (LDA) classifier, first projects the data into the low-dimensional subspace spanned by the eigenvectors of $S_T^{\dagger} S_B$, and then applies the nearest neighbor classifier using cosine similarity measure. LDA baseline is *parameter-free*, and the parameter settings of other methods follow the reference [4].

TABLE 2
Recognition Rates (%) on the Extended Yale B Database

Features	Classifiers	Feature Dimension			
		30	56	120	504
Eigenfaces	NS	89.9	91.1	92.5	93.2
	L-SVM	70.6	84.3	93.1	96.8
	SRC	86.5	91.6	93.9	96.8
	LDA	75.0	91.1	96.1	**99.4**
Laplacianfaces	NS	89.0	90.4	91.9	93.4
	L-SVM	72.0	85.0	94.0	97.7
	SRC	87.5	91.7	93.9	96.5
	LDA	78.5	88.2	95.4	**98.8**
Fisherfaces	NS	81.9	N/A	N/A	N/A
	L-SVM	86.7	N/A	N/A	N/A
	SRC	86.1	N/A	N/A	N/A
	LDA	**98.8**	N/A	N/A	N/A
Randomfaces	NS	87.3	91.5	93.9	94.1
	L-SVM	48.8	68.6	83.4	91.4
	SRC	82.6	91.5	95.5	**98.1**
	LDA	86.9	90.2	91.6	97.3
Downsample	NS	80.8	88.2	91.1	93.4
	L-SVM	48.9	69.5	79.0	91.6
	SRC	74.6	86.2	92.1	**97.1**
	LDA	77.8	86.9	92.5	96.4

Table 2 enumerates the comparative performance of tested classifiers using various feature spaces. LDA baseline achieves recognition rates between 91.6 and 96.1 percent for all 120D feature spaces and a maximum rate of 99.4 percent with 504D Eigenfaces[2]. In contrast, the maximum recognition rate for SRC is only 98.1 percent. In high dimension, such as 504D, the performances of various features in conjunction with both SRC and LDA converge, with conventional features and unconventional features performing similarly. Wright et al. [4] explained this accuracy coverage by the theory of compressive sensing: 504 linear measurements should suffice for sparse recovery in the EYB database. However, even with the Randomfaces that is designate for compressive sensing, the accuracy difference between LDA and SRC is less than one percent (98.1 versus 97.3 percent in 504D space). Moreover, it should be noted that LDA baseline is irrelevant to sparse recovery, but achieves the three highest accuracies, i.e., 99.4, 98.8, 98.8 percent, over the whole experiment. These results clearly suggest that the tested features can achieve high accuracy simply because they are effective to preserve the class separability measured by $\mathrm{Tr}\{S_T^{\dagger} S_B\}$.

An additional evidence to support our claim is that, in the 30D Fisherfaces feature space, LDA baseline achieves 98.8 percent accuracy while SRC only 86.3 percent. This is because that the low-dimensional Fisherfaces features preserve the class separability, but discard most reconstructive information [40]. Clearly, besides the prerequisite discriminatory information, SRC requires enough reconstructive information to assure the coding coefficients meaningful. In this experiment, SRC implicitly takes advantage of the class separability that resides in the reconstructive bases to achieve nearly perfect accuracy.

To ensure the equitable comparison, we have conducted an additional experiment using the Eigenfaces feature as suggested in

2. LDA in 504D eigenspace performs the best among all dimension reduction approaches and among all classifiers. As suggested in [18], PCA helps improve the classification accuracy because it has some roles in removing the unreliable dimension. Specifically, due to the high dimension and small sample size of the face dataset, the components corresponding to small eigenvalues largely deviate from the population variances [37], removing them by PCA not only circumvents the singularity problem of the scatter matrices, but, more importantly, obtains a more reliable estimation of the eigen-spectrum for discriminant analysis [18]. This finding suggests that the accuracy of LDA can be further improved by a selection of PCA features [38] or proper eigen-specturum regularization [39].

TABLE 3
Comparative Recognition Rates of SLRC and Other Recognition Methods

Algorithms	Dictionary Size	Accuracy
ℓ_2 [6]	300×1300	$94.39 \pm 1.35\%$
Nearest Subspace [36]	300×1300	$90.24 \pm 2.16\%$
Random OMP [6]	300×1300	$84.85 \pm 3.43\%$
Hash OMP [6]	300×1300	$86.92 \pm 3.44\%$
CRC [7]	300×1300	$93.76 \pm 0.92\%$
SRC [4]	300×1300	$92.82 \pm 0.95\%$
LDA	300×1300	$96.55 \pm 0.25\%$
ESRC [43]	300×2600	$96.88 \pm 0.71\%$
LR+SI [44]	300×1300	$96.98 \pm 0.81\%$
SDR-SLR [27]	$(41 \times 30) \times 2600$	$98.15 \pm 0.54\%$
SLRC-ℓ_2	300×1400	$97.25 \pm 0.64\%$
SLRC-ℓ_1	300×1400	$\mathbf{98.31 \pm 0.44\%}$

The results of the first four rows are cited from [6] under identical experimental settings.

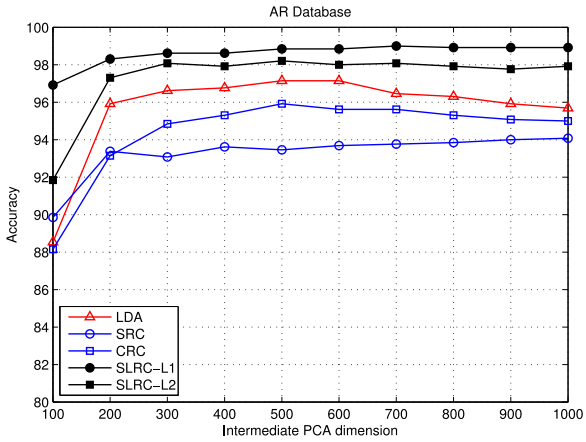


Fig. 4. The recognition performance as the dimension of the intermediate PCA subspace increases on the AR dataset.

[41]. In the training stage, we apply the 10-fold cross validation to select the optimum values of the parameters (subspace dimension d of NS, C of L-SVM, and λ in SRC) at the uniformly sampled intermediate PCA dimensions (p), i.e., 100, 200, ... 1000. We find that the performance of all classifiers become steady when the retained PCA dimension is larger than 500. The best testing accuracy of the NS, L-SVM, SRC, LDA classifiers are 93.6 percent (p=800, d=20), 97.1 percent (p=900, C=1000), 97.2 percent (p=1000, λ=0.001), and 99.8 percent (p=800) respectively. These best accuracies are similar to the ones using 504D eigenspace, which validates our observations on Table 2.

In summary, LDA, SRC, and L-SVM all achieve excellent performance on this controlled dataset with large class separability. One should pay more attention to the general problem with uncontrolled and undersampled training set.

4.3 Recognition with Contaminative Training Set

The AR database consists of over 3,000 frontal images of 126 individuals. There are 26 images of each individual, taken at two different occasions [42]. The faces in AR contain variations such as illumination change, expressions and facial disguises (i.e., sun glasses or scarf). We randomly select 100 subjects (50 male and 50 female) for our experiments, and the images are cropped with dimension 165×120. For each subject, the 26 images are randomly permuted and then the first half is taken for training and the rest for testing. In this way, we have 1,300 training images and 1,300 test images. For statistical stability, 10 different training and test set pairs are generated by randomly permuting, and averaged accuracy and standard deviation are reported. Except the SDR-SLR method that works on the downsampled images, all tested methods are applied on the 300 dimensional PCA space following the setting in [6].

While SRC achieves nearly perfect accuracy on the controlled EYB database, it yields only an average accuracy of 92.82 percent that is notably worse than the 94.39 percent accuracy of basic ℓ_2 approach [6]. As suggested by Wright et al. [45], SRC suffers from the corrupted and occluded training images occlusion that would break the sparsity assumption. In this situation, class-specific concentration of coefficients is violated. For example, the test images wearing sunglasses tend to induce large coding coefficients on the subjects also with sunglasses. ℓ_2 regularization based CRC performs slightly better than SRC. These results are consistent with that found by Shi et al. [6]. Furthermore, LDA baseline outperforms both SRC and CRC by a large margin, which shows that the collaborative representations cannot fully exploit the class separability residing in the uncontrolled training images.

However, one should not deny the usefulness of the CR solely based on the inferiority of the training image based dictionary. We find that *the discrimination power of coding coefficients relies heavily on the suitable choice of dictionary bases*. Specifically, Table 3 fairly compares SRC, Extended SRC (ESRC) [43], Low-rank recovery with structure incoherence (LR+SI) [44], SLRC-ℓ_2 and SLRC-ℓ_1 in the 300 dimensional PCA space. By simply re-construct the dictionary by the class centroids based decomposition, the SLRC-ℓ_2 dramatically boosts the recognition accuracy to about 97 percent. The ESRC method, which appends an intra-class dictionary to the training samples, also increases the accuracy to about 97 percent, but using a much larger dictionary of 2600 bases. When imposed to superposed representation, SLRC-ℓ_1 outperforms SLRC-ℓ_2 by a margin nearly two times standard deviation. Clearly, sparsity constraint is useful in selecting the intra-class variation bases of superposed representation.

The recently proposed SDR-SLR method also achieves similar accuracy to SLRC, in which the supervised low-rank dictionary learning is effective to separate the intra-class variation. Note that the centroids based dictionary of SLRC is parameter-free and very efficiently to construct, and the size of dictionary for classification is similar to CR. In comparison to SDR-SLR, SLRC provides a more simple and efficient solution to the contaminative training set. Its superior accuracy indicates that class centroid indeed provides a stabilized prototype by feature averaging, and the separated contaminative variation can be shared across classes.

Inspired by the previous finding that a good choice of retained PCA greatly improves the LDA performance [14], [15], [46], we investigate how the PCA dimension affects the recognition performance of CR (on the first out of the ten training/test partitions). Fig. 4 confirms that, on this AR dataset with relatively large sample size and controlled variations, the performance of the classifier does not depend so much on the dimension of the intermediate PCA subspace. In this experiment, we can safely select all dimensions of this subspace. Moreover, SLRC displays a steady improvement on the SRC/CRC and LDA in all dimensionality, clearly suggesting that *the proposed superposed representation successfully overlays the advantages of the stability of the CR, and the discrimination ability of the LDA (by the decomposed representation)*.

4.4 Recognition with Small Uncontrolled Training Set

The FRGC version 2.0 is a large-scale face database established under uncontrolled indoor and outdoor settings [47]. We used a subset (316 subjects with no less than ten samples, 7,318 images in total) of the query face dataset, which has large lighting, accessory (e.g., glasses), expression variations and image blur, etc. We randomly chose 3-5 samples per subject as the training set, and used the remaining images for testing. The aligned images are downsampled to 42×32 and the experiments were run 10 times to calculate the mean and standard deviation. Some downsampled images

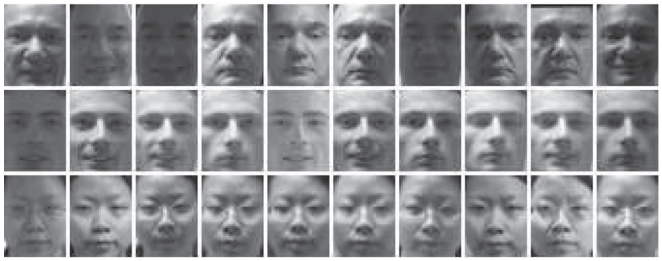


Fig. 5. Some sample images from the FRGC 2.0 database.

are shown in Fig. 5. We compare the proposed SLRC with seven latest dictionary learning based methods including joint dictionary learning (JDL) [21], dictionary learning with commonality and particularity (COPAR) [22], label consistent KSVD (LCKSVD) [23], discriminative KSVD (DKSVD) [24], dictionary learning with structure incoherence (DLSI) [25], Fisher discrimination dictionary learning (FDDL) [26], sparse- and dense-hybrid representation framework with supervised low-rank dictionary (SDR-SLR) [27]. Except the SDR-SLR method that works on the downsampled images, all tested methods are applied on the 300 dimensional PCA space following the setting in [26].

The comparative results are listed in Table 4. SLRC-ℓ_1 performs better than SLRC-ℓ_2 in all the three cases. This suggests the sparsity constraint is crucial for the superposed linear representation. It can be seen that in most cases SLRC-ℓ_1 can have visible improvement over all the other methods. SLRC outperforms SRC by about 8.6, 6.6, and 5.1 percent when there are 3, 4, and 5 training images per person. This clearly shows the effectiveness of the "naive" parameter-free decomposition of the centroid and intra-class variation in SLRC, especially on the small sample size cases. By this simple decomposition, SLRC also outperforms seven state-of-the-art dictionary learning methods with sophisticated settings.

The COPAR method [22] also considers the common and particular components in the dictionary, but the learned dictionary cannot extract the accurate inter/intra-class components as indicated by its inferior accuracy. The SDR-SLR method [27], which develops a class-wise supervised low rank decomposition to learn the intra-class dictionary, achieves comparable accuracy that is better than SLRC-ℓ_2 but slightly worse than SLRC-ℓ_1. This may be because that the dense coefficients of SDR-SLR are not optimal for a dictionary of uncontrolled overcomplete intra-class variation bases. Compared with these sophisticated dictionary learning methods, SLRC indeed provides a simple but powerful solution to generalize the CR to the uncontrolled face recognition problem.

As in the AR experiment, we investigate how the choice of retained PCA dimension affects the recognition performance (on the first out of the ten training/test partitions). Different from the observation on AR dataset, Fig. 6 shows that LDA achieves comparable performance with SRC/CRC within the 200-400 dimension, but deteriorates dramatically as the dimensionality becomes higher. This observation on curse of dimensionality is consistent with previous studies [15] on the small size dataset. This is because the trivial components are enlarged by the whitening process of LDA, and these components tend to be noisy when the training samples are insufficient. In contrast, even on this small size dataset with uncontrolled variations, the four CR methods achieve steady accuracy across varying retained PCA dimensions. Their accuracies steadily increase when the dimension is larger than 300. Identical to the results on the AR database, SLRC inherits the stability of CR and displays a steady improvement on the SRC/CRC and LDA in all dimensionality.

TABLE 4
The Face Recognition Rates (%) of Competing Methods on the FRGC 2.0 Database with N Training Samples per Person

Algorithms	$N = 3$	$N = 4$	$N = 5$
SRC [4]	80.4 ± 0.6	87.0 ± 0.6	87.7 ± 0.4
CRC [7]	82.6 ± 0.6	87.4 ± 0.6	89.7 ± 0.3
LDA	79.0 ± 0.8	87.0 ± 0.7	90.2 ± 0.7
NSC [36]	54.7 ± 0.7	63.0 ± 0.6	69.3 ± 0.6
SVM [35]	57.1 ± 0.7	66.2 ± 0.7	72.9 ± 0.7
DKSVD [24]	72.2 ± 0.6	77.2 ± 0.7	79.7 ± 0.7
LCKSVD [23]	75.7 ± 0.6	78.1 ± 0.5	79.8 ± 0.8
DLSI [25]	86.7 ± 0.6	91.4 ± 0.5	93.5 ± 0.3
COPAR [22]	81.3 ± 0.6	86.9 ± 0.6	89.5 ± 0.6
JDL [21]	83.0 ± 0.7	88.2 ± 0.5	91.2 ± 0.5
SDR-SLR [27]	89.5 ± 0.8	93.1 ± 0.4	94.2 ± 0.4
FDDL [26]	89.0 ± 0.8	92.9 ± 0.3	95.1 ± 0.3
SLRC-ℓ_2	85.0 ± 0.6	91.1 ± 0.5	92.8 ± 0.3
SLRC-ℓ_1	**90.0** ± 0.7	**93.6** ± 0.6	**95.2** ± 0.4

4.5 Recognition with Single Sample Per Person

The final experiment aims to evaluate *the applicability of SLRC with only a single training sample per person*. The experiment follows the standard data partitions of the FERET database [48]. The images are first normalized by a similarity transformation that sets the centers of the eyes at the settled coordinates. Fig. 7a shows some cropped images which are used in our experiments. Note that the images of FERET database contain complex intra-class variability, since they are acquired in multiple sessions during several years. As there is only a single sample per gallery class, we construct the intra-class variation matrix from the standard training image set of the FRGC Version 2 database [47], which contains 12,766 frontal images of 222 people taken in the uncontrolled conditions. Fig. 7b shows some intra-class differences computed by (13) from this image set. Note that the collection of the FRGC database is totally independent from the FERET database. Hence, in this experiment, the variation dictionary is required to universally represent the complex facial variations under uncontrolled conditions.

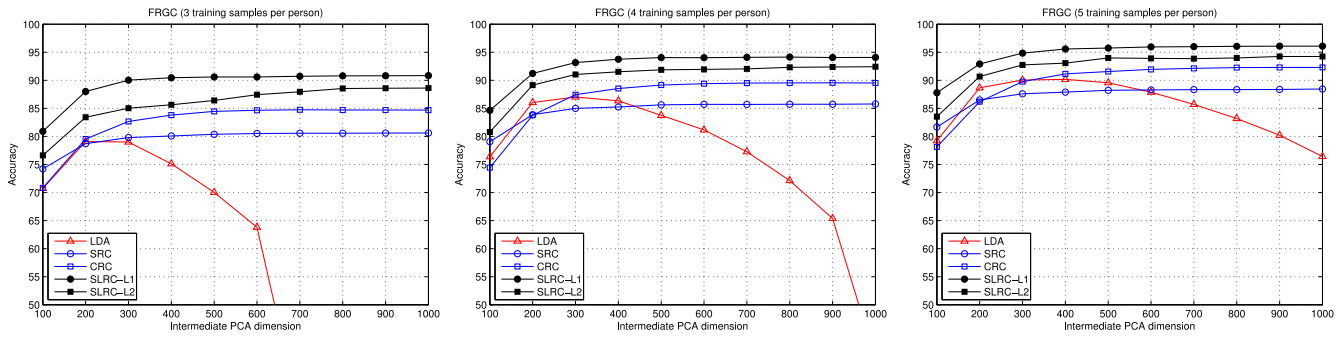


Fig. 6. The recognition performance as the dimension of the intermediate PCA subspace increases on the FRGC dataset using (a) 3 (b) 4 (c) 5 training samples per person.

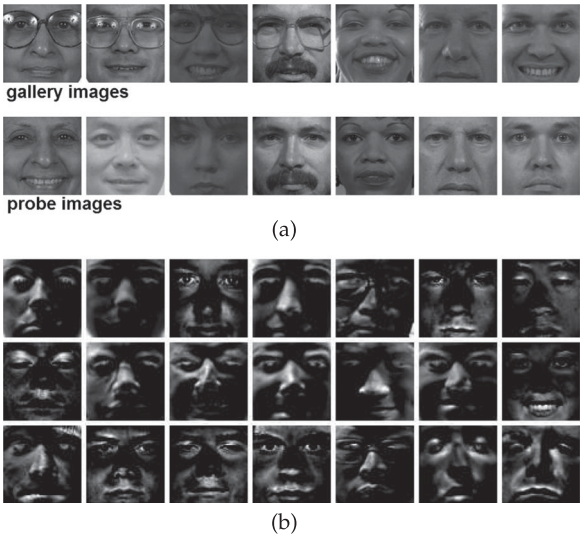


Fig. 7. (a) The cropped images of some gallery images and corresponding probe images in the FERET database. (b) Example images of the differences to the class centroid computed from the FRGC version 2 database.

As the performance of SLRC increases stably with higher the PCA dimension, we select the PCA dimension as high as 1,000, and investigate the regularization effects on the uncontrolled and over-complete variation dictionary. Specifically, we first test the performance of the SLRC-ℓ_2 by increasing the parameter λ from 0.000001 to 100, as shown in Fig. 8. When the value of λ is relatively large in the range of $[0.1, 10]$, ℓ_2-norm regularization obtains its optimal performance. However, the optimal performance of ℓ_2-norm regularization is significantly lower than that of SLRC-ℓ_1 tested with limited number of $\lambda = \{0.0005, 0.005, 0.01\}$. The superiority of SLRC-ℓ_1 seems more apparent on the dup1 and dup2 set. A large margin over 10 percent accuracy is observed on dup1 set when comparing SLRC-ℓ_1 with SLRC-ℓ_2. This implies that sparse coefficients indeed play a crucial role in face recognition given an uncontrolled and over-complete dictionary.

For comprehensive results, we also extract the Gabor feature, LBP feature and PCANet [50] feature for classification besides the pixel intensity. For each feature, we test the recognition performance in the reduced PCA dimension of 1000. In total, there are 16 test cases (4 probes×4 features) and Table 5 lists the comparative performance between SRC and SLRC in all cases. Although the variation dictionary is constructed from the FRGC database, SLRC improves the recognition rates on the FERET database in all the 16 test cases, indicating that the intra-class variability of face is sharable

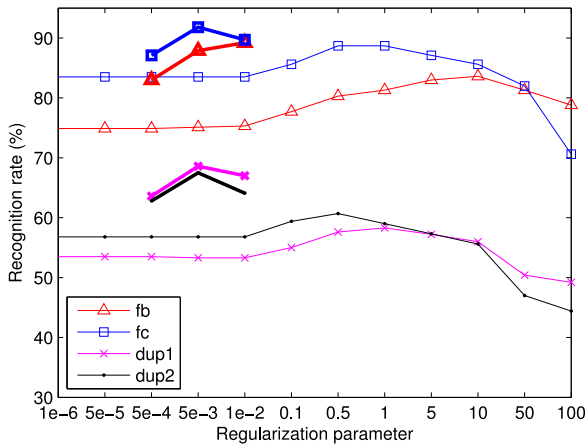


Fig. 8. The recognition rates of SLRC with ℓ_1-regularization (plotted by thick symbols) and ℓ_2-regularization (plotted by the thin symbols) as a function of the value of λ.

TABLE 5
Comparative Recognition Rates of SRC and SLRC on FERET Database Using Single Training Sample per Person

Features	Methods	fb	fc	dup1	dup2
Intensity	SRC	85.2	76.3	63.9	57.3
	SLRC-ℓ_1	87.9	91.8	68.6	67.5
Gabor [14]	SRC	93.0	97.4	73.0	78.6
	SLRC-ℓ_1	96.7	99.5	80.7	85.5
LBP [49]	SRC	96.9	93.8	87.7	85.0
	SLRC-ℓ_1	98.0	99.5	90.6	90.2
PCANet [50]	SRC	98.8	99.0	94.9	92.3
	SLRC-ℓ_1	99.4	100.0	**96.3**	**95.7**

even when the generic data are collected from different conditions and camera set-ups. Our results also suggest that the superposed linear representation model is feasible for various feature representations, and thus it can be integrated with more informative features to address uncontrolled face recognition problem. When applied on the PCANet feature, SLRC achieves state-of-the-art performance on FERET database with a single training sample per person. The improvement is visible on the dup1 and dup2 probe sets, which indicates the sparse coding can play an important role on selecting bases to represent the real-world age variation.

5 Conclusions

The experiments suggest a number of conclusions:

1. The class superability of the controlled face dataset, such as Extended Yale B database, is sufficiently large. Both the traditional baseline algorithms that characterize the quantity of $\mathrm{Tr}\{S_T^{\dagger}S_B\}$, and the collaborative representation methods, such as SSC and SRC, can achieve excellent clustering and classification performance. The research should pay more attention to the general problem with uncontrolled and undersampled datasets.
2. By the "naive" centroid based dictionary decomposition, the new SLRC successfully overlays the advantages of the robustness of the collaborative representation, and the discrimination ability of the LDA (by the decomposed representation).
3. Although the class centroid is generally an approximated representation, in practice SLRC competes well with more sophisticated dictionary learning techniques in our experiments. Moreover, SLRC does not substantially increase computation and storage compared to basic collaborative representation methods such as SRC and CRC.
4. By constructing intra-class dictionary from the generic dataset, SLRC is effective to address the recognition problem with single sample per person. Thanks to its simple representation assumption, it is also applicable to various feature descriptors besides pixel intensity, by which state-of-the-art face recognition performance can be achieved.
5. When the variation dictionary is overcomplete, sparse coefficient regularizer plays a crucial role on recognition: SLRC with ℓ_1-sparsity lasso solution outperforms ℓ_2 ridge regression solution by a large margin for face recognition.

Acknowledgments

The authors would like to thank the anonymous reviewers for their thoughtful and constructive remarks that are helpful to improve the quality of this paper. This work was partially supported by the National Natural Science Foundation of China under Grants 61573068, 61471048, 61375031, and 61532006, Beijing Nova Program under Grant No. Z161100004916088.

REFERENCES

[1] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[3] W. Deng, J. Hu, J. Lu, and J. Guo, "Transform-invariant PCA: A unified approach to fully automatic facealignment, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1275–1284, Jun. 2014.

[4] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[5] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[6] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 553–560.

[7] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 471–478.

[8] W. Deng, J. Hu, and J. Guo, "In defense of sparsity based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 399–406.

[9] R. Penrose, "A generalized inverse for matrices," in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge, U.K.: Cambridge Univ Press, 1955, vol. 51, pp. 406–413.

[10] W. Deng, J. Hu, X. Zhou, and J. Guo, "Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning," *Pattern Recognit.*, vol. 47, no. 12, pp. 3738–3749, 2014.

[11] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[12] A. M. Martinez and M. Zhu, "Where are linear feature extraction methods applicable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1934–1944, Dec. 2005.

[13] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenfaces: Probabilistic matching for face recognition," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 30–35.

[14] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[15] J. Bekios-Calfa , J. M. Buenaposada, and L. Baumela, "Revisiting linear discriminant techniques in gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 858–864, Apr. 2011.

[16] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.

[17] P. Howland and H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 8, pp. 995–1006, Aug. 2004.

[18] X. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 16–26, Mar. 2011.

[19] R. Jenkins and A. Burton, "100 percent accuracy in automatic face recognition," *Sci.*, vol. 319, no. 5862, pp. 435–435, 2008.

[20] W. Deng, J. Guo, J. Hu, and H. Zhang, "Comment on 100 percent accuracy in automatic face recognition," *Sci.*, vol. 321, no. 5891, 2008, Art. no. 912.

[21] N. Zhou and J. Fan, "Jointly learning visually correlated dictionaries for large-scale visual recognition applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 715–730, Apr. 2014.

[22] S. Kong and D. Wang, "A dictionary learning approach for classification: separating the particularity and the commonality," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 186–199.

[23] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

[24] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2691–2698.

[25] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3501–3508.

[26] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, 2014.

[27] X. Jiang and J. Lai, "Sparse and dense hybrid representation via dictionary decomposition for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1067–1079, May 2015.

[28] W. Deng, J. Hu, and J. Guo, "Gabor-eigen-whiten-cosine: A robust scheme for face recognition," in *Proc. Int. Workshop Analysis and Modeling of Faces and Gestures*. Berlin, Germany: Springer, 2005, pp. 336–349.

[29] W. Deng, J. Hu, J. Guo, W. Cai, and D. Feng, "Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach," *Pattern Recognit.*, vol. 43, no. 5, pp. 1748–1762, 2010.

[30] W. Deng, Y. Liu, J. Hu, and J. Guo, "The small sample size problem of ICA: A comparative study and analysis," *Pattern Recognit.*, vol. 45, no. 12, pp. 4438–4450, 2012.

[31] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2006, pp. 94–106.

[32] G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *Int. J. Comput. Vis.*, vol. 81, no. 3, pp. 317–330, 2009.

[33] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1801–1807.

[34] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 29th Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[35] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, Art. no. 27.

[36] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[37] X. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 931–937, May 2009.

[38] M. Zhu and A. M. Martinez, "Selecting principal components in a two-stage LDA algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 1, pp. 132–137.

[39] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 383–394, Mar. 2008.

[40] S. Fidler, D. Skočaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 337–350, Mar. 2006.

[41] P. Dago-Casas, D. González-Jiménez, L. L. Yu, and J. L. Alba-Castro, "Single-and cross-database benchmarks for gender classification under unconstrained settings," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2152–2159.

[42] A. M. Martinez and R. Benavente, "The AR face database," Universitat Autònoma de Barcelona, Barcelona, Spain, CVC Tech. Rep. #24, Jun. 1998.

[43] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.

[44] C. Chen, C. Wei, and Y. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2618–2625.

[45] J. Wright, A. Ganesh, A. Yang, Z. Zhou, and Y. Ma, "Sparsity and robustness in face recognition," *arXiv:1111.1014*, 2011.

[46] W. Deng, J. Hu, J. Guo, W. Cai, and D. Feng, "Emulating biological strategies for uncontrolled face recognition," *Pattern Recognit.*, vol. 43, no. 6, pp. 2210–2223, 2010.

[47] P. J. Phillips, et al., "Overview of the face recognition grand challenge," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 947–954.

[48] P. J. Phillips, H. Moon, P. Rizvi, and P. Rauss, "The feret evaluation method for face recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 0162–8828, Oct. 2000.

[49] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[50] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015.