Sparse And Dense Hybrid Representation via Dictionary Decomposition for Face Recognition

Xudong Jiang, *Senior Member, IEEE,* and Jian Lai, *Student Member, IEEE*

Abstract—Sparse representation provides an effective tool for classification under the conditions that every class has sufficient representative training samples and the training data are uncorrupted. These conditions may not hold true in many practical applications. Face identification is an example where we have a large number of identities but sufficient representative and uncorrupted training images cannot be guaranteed for every identity. A violation of the two conditions leads to a poor performance of the sparse representation-based classification (SRC). This paper addresses this critic issue by analyzing the merits and limitations of SRC. A sparse- and dense-hybrid representation (SDR) framework is proposed in this paper to alleviate the problems of SRC. We further propose a procedure of supervised low-rank (SLR) dictionary decomposition to facilitate the proposed SDR framework. In addition, the problem of the corrupted training data is also alleviated by the proposed SLR dictionary decomposition. The application of the proposed SDR-SLR approach in face recognition verifies its effectiveness and advancement to the field. Extensive experiments on benchmark face databases demonstrate that it consistently outperforms the state-of-the-art sparse representation based approaches and the performance gains are significant in most cases.

Index Terms—Sparse representation, classification, dictionary learning, low-rank matrix recovery, face recognition

1 Introduction

High data dimensionality and lack of human knowledge about the effective features to classify the data are two challenging problems in computer vision and pattern recognition. Face recognition remains a hot research topic after extensive research in the past two decades not only due to its huge application potential. It provides a good test bed to show how the two key computer vision problems are solvable because large and unambiguous test databases are available for face recognition problem. Holistic approach, or called appearance based approach, which applies machine learning techniques on the whole image for the both feature extraction and classification, provides a plausible tool to tackle these two difficult computer vision problems.

For the machine learning based feature extraction or dimensionality reduction, representative approaches include the principal component analysis [1], [2], the linear discriminant analysis [3], [4], probabilistic subspace learning [5], [6] and their extensions to the locality preservation [7], [8] and in the kernel space [9], [10]. Some of them are successful in reducing the data to a very low dimensional subspace yet keeping the same or even enhancing the classification performance. The principle and rationale behind the classification performance enhancement by dimensionality reduction are analyzed in [11].

Comparing to the machine learning based feature extraction or dimensionality reduction, research on the classification had not been very active in this area before the sparse representation based classification was introduced a few years ago. Most approaches simply apply the nearest neighborhood (NN) classifier or the minimum Mahalanobis distance classifier in a subspace obtained by a dimensionality reduction procedure. Extensions of the NN classifier in the high dimensional space include the nearest feature line [12], the nearest plane [13] and the nearest subspace [13]–[15] classifiers. Although the linear regression classifier (LRC) [16], [17] represents the query image by a linear combination of the class-specific training samples, it is not difficult to see that it is in fact equivalent to the nearest subspace [13]–[15] classifiers. The commonalities of all these classifiers are that they are not robust to the heavy image corruption caused by the outlier pixels and occlusions and that they evaluate the relation between the query image and the training samples of each individual class one by one separately.

Different from the above classifiers, the sparse representation [18]–[20] of the query image by training samples of all classes is applied to design the classifier for face recognition [21]. The sparse representation-based classifier (SRC) [21], in our opinion, significantly differentiates itself from the above classifiers in three aspects. One is the utilization of training samples of all classes collaboratively to represent the query images and another is the sparse representation code that coincides with the general classification target. The last is the ℓ_1-norm minimization of the representation error that enables SRC to recognize query images heavily corrupted by outlier pixels and occlusions. These three merits of SRC lead to some encouraging and impressive face recognition results, which attract great interest in further research on SRC. Many extensions of SRC are proposed in recent years, such as Gabor feature based

• *X.D. Jiang and J. Lai are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Nanyang Link, Singapore 639798.*
E-mail: exdjiang@ntu.edu.sg, jlai1@ntu.edu.sg

SRC [22], SRC with nonnegative constraint [23], Locality-constraint SRC [24], Gaussian kernel error term [25], modular weighted global SRC [26], pose alignment with SRC [27] and regularized robust coding [28].

These extensions, however, have not solved the two fundamental constraints of SRC. One is the carefully controlled training images and the other is the sufficient training samples of each class. A violation of these two conditions results in poor performance of the sparse representation-based classification [27], [29]. The first constrain makes SRC based approaches do not perform well for the corrupted training data caused by outlier pixels and occlusions. The second constraint limits their applications to large scale identification problems where the training data contains large number of identities but sufficient representative images for every identity cannot be guaranteed. Therefore, it is not a surprise that, despite of the impressive results of SRC and many of its variants and extensions developed, a number of works [30]–[33] show questions and doubts about the effectiveness of SRC for image classification.

The both fundamental constraints of the sparse representation-based classification are related to the training data. As SRC directly applies training images as the dictionary for the sparse representation, questions may arise if dictionary learning [34] can help alleviate these two problems of SRC. The objective of the general dictionary learning [35], [36] is to find different optimal dictionaries in representing different target data. The atoms in a dictionary computed from the training database in general capture the most important constitutive component of the target data. However, there is no good reason why the set of atoms is also best for differentiating different classes in the target data. The discriminative dictionary learning [37]–[40] makes a compromise between the data representation and the class separation in finding the optimal atoms of dictionary. This, however, also does not solve the problem of SRC when not all classes have sufficient representative training samples. In addition, all these dictionary learning approaches do not work well if the training images are heavily corrupted by the outlier pixels and occlusions [34]. Instead of the conventional dictionary learning approaches, it seems that we need separate the image corruptions of the training samples from the dictionary.

To overcome the first constraints of SRC – corrupted training data, low-rank matrix recovery, also called robust PCA (RPCA) [41], provides a tool to separate outlier pixels and occlusions from the training images. A sub-dictionary is learnt from each class separately by decomposing the training data of each class into a low-rank matrix and a sparse matrix in [42], [43]. However, optimizing sub-dictionaries to be low-rank for each class might reduce the diversity of atoms within the sub-dictionary and hence might decrease the dictionary's representation power [44]. Following the idea of subspace clustering by low-rank representation [45], low-rank representation (coefficient matrix of the dictionary) is learnt for image classification [44]. However, the natural cluster structure of the training data may not coincide with the class (identity) structure for a face identification problem due to the small differences between the face identities and the large variations of face images of the same identity. As a result, in seeking a low rank dictionary or a low rank representation, these methods may undesirably remove some components of face identity and/or some face variations needed to represent the query images.

There are some attempts to alleviate the second problem of SRC – lack of variations in the training images for some classes. The extended SRC (ESRC) [46] creates an intra-class variation matrix by subtracting a natural or prototype or centroid image of each class from training images of the same class. The variation matrix is then appended to the raw training data matrix as the second part of the dictionary. The classification only utilizes the representation coefficients of the first part. Although the query images can be better represented by some variations of other classes in the second part of the dictionary, the same variations are also contained in the first part. Therefore, the introduced second part of the dictionary may not well solve the problem of the first part. This leads the same authors of ESRC to further propose a superposed SRC (SSRC) [47] by replacing the first part with a natural or prototype or centroid image of each class. However, this brings another problem that the identity of the query image is represented and determined by only a single image, which may result in unstable or unreliable classification.

This work aims to solve the two fundamental limitations of SRC – the lack of variations in the training images for some classes and the corrupted training data. We propose a hybrid representation of the query image by a sparse combination of a class-specific dictionary and a dense combination of a common intra-class variation dictionary. Thus, a query image is better represented with the collaboration of image variations of other classes, which leads the sparse representation part better to coincide with the specified class membership. Towards this end, we further propose a procedure to decompose the raw training data into a class-specific dictionary, a common intra-class variation dictionary and a sparse corruption matrix. In this way, the corruption matrix will also be better separated from two lower-rank matrices than from the higher-rank mixture of them. In implementation, the proposed approach needs solve a nuclear norm regularized optimization, which is a convex minimization problem solvable in polynomial time.

2 Sparse and Dense Hybrid Representation

An image of m pixels is arranged in a column vector. Let $\mathbf{D}_i \in \Re^{m \times n_i}$ stack n_i m-dimensional training samples of the ith class. An unlabeled query image $\mathbf{y} \in \Re^m$ is represented or approximated by a linear combination of the training samples from a class as

$$\mathbf{y} = \mathbf{D}_i \boldsymbol{\alpha}_i + \mathbf{e}_i, \quad i = 1, 2, ..., c, \tag{1}$$

where, $\boldsymbol{\alpha}_i$ is a coefficient vector associated with the training samples of class i and $\mathbf{e}_i$ is the approximation error. It is widely assumed that images of a specific class lie in a linear subspace [3], [48]. Thus, the linear regression

classifier (LRC) [16], [17] assigns the class label of $\mathbf{y}$ to class i that produces the smallest error $\|\mathbf{e}_i\|_2$. This is in fact equivalent to a nearest subspace [13]–[15] classifier as $\|\mathbf{e}_i\|_2$ is the distance from $\mathbf{y}$ to the subspace spanned by columns of $\mathbf{D}_i$.

There is a problem if $n_i << m$ or some classes do not have representative training samples. The variation between the query image and the training samples of the same class could be larger than those of some other classes, which results in misclassification. Even for a large number of training samples of each class, there is another problem that errors of some classes could be so small that the classification based on them is unreliable. This can be understood by the fact that there exists a representation $\boldsymbol{\alpha}_i$ leading to $\mathbf{e}_i = 0$ for any class as long as its training samples are not linear dependent and $n_i \geq m$.

Now consider a collaborative representation of the query image by training samples of all classes as

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{e}, \tag{2}$$

where $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_c] \in \Re^{m \times n}$ stacks training samples of all c classes, $n = \sum_{i=1}^{c} n_i$. If $\text{rank}(\mathbf{D}) = m = n$, we can have a unique perfect representation $\boldsymbol{\alpha}$ that leads to $\mathbf{e} = 0$ for any query image $\mathbf{y}$. If the dictionary $\mathbf{D}$ is over-complete, i.e. $n > m$, the perfect representation $\boldsymbol{\alpha}$ is not unique, i.e. there are an infinite number of solutions $\boldsymbol{\alpha}$ that lead to $\mathbf{e} = 0$. Thus, a perfect representation with the sparsity constraint on $\boldsymbol{\alpha}$ can be found. We can achieve a more sparse solution $\boldsymbol{\alpha}$, or get it even on an under-complete dictionary, $n < m$, by relaxing the requirement for the perfect representation. Therefore, we look for a sparse linear representation $\boldsymbol{\alpha}$ of $\mathbf{y}$ over $\mathbf{D}$ with a bounded energy of the representation error $\|\mathbf{e}\|_2^2 < \varepsilon$.

The sparsity of $\boldsymbol{\alpha}$ is measured by the number of its nonzero elements, i.e. ℓ_0-norm of $\boldsymbol{\alpha}$, $\|\boldsymbol{\alpha}\|_0$. This optimization problem is formulated as

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0, \text{ s.t. } \|\mathbf{e}\|_2^2 < \varepsilon, \text{ where } \mathbf{e} = \mathbf{y} - \mathbf{D}\boldsymbol{\alpha}. \tag{3}$$

Although this minimization problem is NP-hard, the compressed sensing theory reveals that, if its solution is sufficient sparse, it is equivalent to the following convex relaxation ℓ_1-norm minimization [49]

$$\min_{\boldsymbol{\alpha},\mathbf{e}} \|\boldsymbol{\alpha}\|_1 + \beta\|\mathbf{e}\|_2^2, \text{ where } \mathbf{e} = \mathbf{y} - \mathbf{D}\boldsymbol{\alpha}, \tag{4}$$

where β is a constant for a compromise between the sparsity of $\boldsymbol{\alpha}$ and the representation error $\|\mathbf{e}\|_2^2$.

To enable a robust representation of heavily corrupted query image by outlier pixels and occlusions, SRC [21], [29] replaces the ℓ_2-norm of the representation error by ℓ_1-norm in the optimization

$$\min_{\boldsymbol{\alpha},\mathbf{e}} \|\boldsymbol{\alpha}\|_1 + \beta\|\mathbf{e}\|_1, \text{ s.t. } \mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{e}. \tag{5}$$

From (2) and (5) we see three merits of SRC: the collaborative representation using training samples of all classes $\mathbf{D}$; the sparse representation code $\boldsymbol{\alpha}$; and the ℓ_1-norm minimization of the representation error $\|\mathbf{e}\|_1$. These merits make the sparse representation based approaches a great success in image restoration [50], [51] and face recognition [21], [27], [29] in some scenarios. However, SRC does not perform well for corrupted training data or if the sufficient representative training samples are not provided for every class. Thus, a number of works [30]–[33] show questions and doubts about the effectiveness of sparse representation for image classification.

We argue that the sparse representation directly coincides with the general objective of the classification because the desired output or target of a classification system is 1 for correct class and 0 for all others, which is exactly a sparse code of the query image. In addition, the collaborative representation compensates the limited number of representative samples of a single class, though only partially due to the sparse constraint. The optimization of the sparse coefficients lets every training sample compete against the others to win its representation share of the query image. This is a good discriminating process.

Problem is that images contain not only the identity information but also much other information such as age, gender, ethnic, expression and illumination. Why must the significant coefficients in $\boldsymbol{\alpha}$ be won by the samples of the same identity as the query image, not of the same expression or the same illumination? Neither (2) nor (5) receives any information about the class label assignment of the training data that specifies the particular classification task. If we can separate the class-specific information from others, the sparse representation of the former will fully coincides with the particular classification target specified by the class labels of the training data.

Therefore, we propose to decompose an image $\mathbf{y}$ into three components as

$$\mathbf{y} = \mathbf{a} + \mathbf{b} + \mathbf{s}, \tag{6}$$

where $\mathbf{a}$ is the class-specific component, $\mathbf{b}$ the non-class-specific variations and $\mathbf{s}$ contains random sparse noise or image corruption. If a dictionary $\mathbf{A}$ that only contains class-specific component of the image is available, we can apply SRC as

$$\mathbf{a} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{e}_a. \tag{7}$$

The sparse vector $\boldsymbol{\alpha}$ of (7) will directly coincide with the class label vector of the query image $\mathbf{y}$ while that of (2) only coincides with the natural image clusters as $\mathbf{a}$ and $\mathbf{A}$ only have the class-specific components but $\mathbf{y}$ and $\mathbf{D}$ contain all other image information.

Although it might be possible to separate the class-specific component $\mathbf{a}$ from $\mathbf{y}$ based on the human knowledge in some applications, it is very difficult if not impossible for many computer vision tasks such as face recognition. Given a labeled training database $\mathbf{D}$, it is possible to decompose it into a class-specific dictionary $\mathbf{A}$, a non-class-specific dictionary $\mathbf{B}$ and a random sparse noise $\mathbf{E}$ based on machine learning. This will be presented in the next section.

However, it is very difficult if not impossible to separate the class-specific component $\mathbf{a}$ from a single unknown query image $\mathbf{y}$. To circumvent this problem, we also represent the non-class-specific component $\mathbf{b}$ by the dictionary $\mathbf{B}$ as

$$\mathbf{b} = \mathbf{B}\mathbf{x} + \mathbf{e}_b. \tag{8}$$

The only purpose of the representation (8) is to make the representation error $\mathbf{e_b}$ as small as possible. Therefore, it is

not necessary to put the sparse constraint on $\mathbf{x}$ and hence (8) should be in general a dense representation.

The summation of (7), (8) and $\mathbf{s}$ yields the proposed sparse- and dense-hybrid representation (SDR) of the query image

$$\mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{9}$$

where $\mathbf{e} = \mathbf{e}_a + \mathbf{e}_b + \mathbf{s}$ is the combined representation error. Fig. 1 illustrates a face image projected to the identity subspace (spanned by $\mathbf{A}$), the common variation subspace (spanned by $\mathbf{B}$) and its sparse residual.

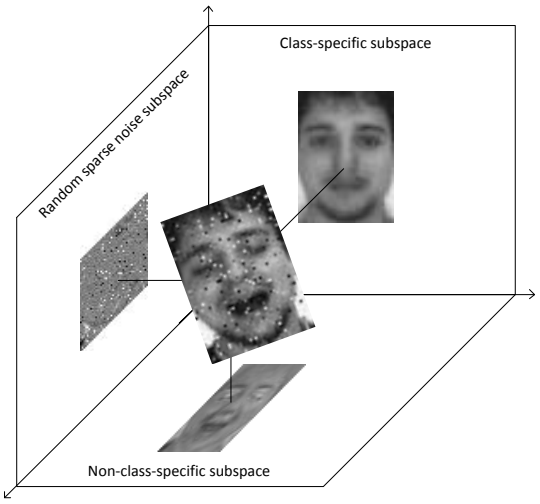


Fig. 1. Image decomposition: a face image projected to the class-specific subspace, the non-class-specific subspace and their residual random sparse noise.

To be robust to the heavily corrupted query image by outlier pixels and occlusions, ℓ_1-norm minimization of the representation error $\mathbf{e}$ is applied. The solution of the proposed SDR, $\boldsymbol{\alpha}$, $\mathbf{x}$ and $\mathbf{e}$ is obtained by solving the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\alpha},\mathbf{x},\mathbf{e}} \|\boldsymbol{\alpha}\|_1 + \gamma\|\mathbf{x}\|_2^2 + \beta\|\mathbf{e}\|_1 \\ &\text{s.t. } \mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{B}\mathbf{x} + \mathbf{e}. \end{aligned} \tag{10}$$

The optimization problem (10) can be solved by the Augmented Lagrange Multiplier (ALM) scheme [52]. The ALM function for (10) is derived as:

$$\begin{aligned} \|\boldsymbol{\alpha}\|_1 + \gamma\|\mathbf{x}\|_2^2 + \beta\|\mathbf{e}\|_1 + \frac{\xi}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{B}\mathbf{x} - \mathbf{e}\|_2^2 \\ + \boldsymbol{\phi}^T(\mathbf{y} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{B}\mathbf{x} - \mathbf{e}), \end{aligned} \tag{11}$$

where $\boldsymbol{\phi}$ is a vector of Lagrange multipliers and ξ a penalty parameter. Algorithm 1 summarizes the solution to problem (10). The subproblem for $\mathbf{e}$ in step 1 can be solved by the soft-thresholding operator [53], the result of $\boldsymbol{\alpha}$ can be achieved by l_1-norm minimization and $\mathbf{x}$ has a closed-form solution.

Let $\mathbf{L}_i \in \Re^{n\times n}$ be a class-label matrix of the training data $\mathbf{D}$ for class i, its element $\mathbf{L}_i(k,k) = 1$ if the kth training sample (the kth column of $\mathbf{D}$) originates from class i and all other elements of $\mathbf{L}_i$ are zero. The representation of the query image $\mathbf{y}$ by the class-specific component of class i and the non-class-specific component of all classes collaboratively is then

$$\mathbf{y} = \mathbf{A}\mathbf{L}_i\boldsymbol{\alpha} + \mathbf{B}\mathbf{x} + \mathbf{e}_i. \tag{12}$$

As discussed earlier, the sparse vector $\boldsymbol{\alpha}$ coincides with the class-label vector (the sum of $\mathbf{L}_i$ over all columns). Therefore, the representation error $\mathbf{e}_i$ will be the closest to $\mathbf{e}$ if the query image $\mathbf{y}$ originates from class i. The class-wise representation residual is defined by

$$r_i(\mathbf{y}) = \|\mathbf{e} - \mathbf{e}_i\|_2 = \|\mathbf{A}(\mathbf{I} - \mathbf{L}_i)\boldsymbol{\alpha}\|_2, \tag{13}$$

where $\mathbf{I}$ is an identity matrix. The query image $\mathbf{y}$ is classified into the class that produces the minimum residual $r_i(\mathbf{y})$.

Sparse representation (2) coincides with the general objective of classification: degenerate all query images to only two different states: one for the correct class and the other for all other classes. However, the competition, in which a training sample gains its share to represent the query image by (2) and (5), is unsupervised. As a result, all information contained in the query image and training samples participates in this competition. Thus, there is no reason why the nonzero significant coefficients should coincide with a specific one of many possible class arrangements that (2) and (5) do not know. For many computer vision tasks, such as face recognition, the class-specified information is only a very small portion of the whole information carried by an image. Therefore, SRC can only work well if all non-class-specific information is nulled out by sufficient representative training samples for every class that enclose all non-class-specific information. This problem is visible in the first row of Fig. 2, which shows an example of SRC where the top 3 significant coefficients are not for the training samples of the same identity as the query image.

Algorithm 1: Solving Problem(10) by Inexact ALM

Input: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{y}$, parameter β and γ.
Initialize: $\boldsymbol{\alpha} = 0$, $\mathbf{x} = 0$, $\mathbf{e} = 0$, $\boldsymbol{\phi} = 0$, $\xi = 1$, $\xi_{max} = 10^6$, $\rho = 1.5$, and $\epsilon = 10^{-6}$.
while *not converged* **do**

1. fix the others and update $\mathbf{e}$ by
$\mathbf{e} = \arg\min \frac{\beta}{\xi}\|\mathbf{e}\|_1 + \frac{1}{2}\|\mathbf{e} - (\mathbf{y} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{B}\mathbf{x} + \frac{1}{\xi}\boldsymbol{\phi})\|_2^2$
2. fix the others and update $\boldsymbol{\alpha}$ by
$\boldsymbol{\alpha} = \arg\min \|\boldsymbol{\alpha}\|_1 + \frac{\xi}{2}\|\mathbf{A}\boldsymbol{\alpha} - (\mathbf{y} - \mathbf{B}\mathbf{x} - \mathbf{e} + \frac{1}{\xi}\boldsymbol{\phi})\|_2^2$
3. fix the others and update $\mathbf{x}$ by
$\mathbf{x} = \xi(2\gamma\mathbf{I} + \xi\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{y} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{e} + \frac{1}{\xi}\boldsymbol{\phi})$
4. update the multipliers
$\boldsymbol{\phi} = \boldsymbol{\phi} + \xi(\mathbf{y} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{B}\mathbf{x} - \mathbf{e})$
5. update ξ by $\xi = min(\rho\xi, \xi_{max})$
6. check the convergence conditions:
$\|\mathbf{y} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{B}\mathbf{x} - \mathbf{e}\|_2^2 < \epsilon$

end
Output: $\boldsymbol{\alpha}$, $\mathbf{x}$ and $\mathbf{e}$

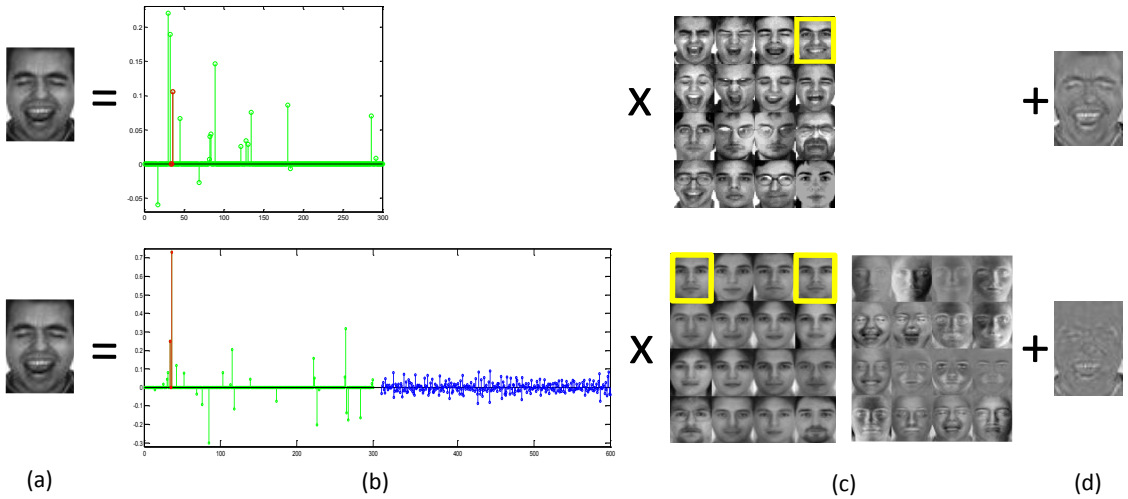


Fig. 2. Comparison of SRC (1^{th} row) and the proposed SDR (2^{nd} row) on AR database $\mathbf{D}$ of 100 persons with 3 images per person: (a) query image; (b) SRC coefficients of $\mathbf{D}$ and SDR coefficients of $\mathbf{A}$ and $\mathbf{B}$; (c) top 16 images in $\mathbf{D}$ of SRC and in $\mathbf{A}$ and $\mathbf{B}$ of SDR; (d) residuals. Sparse coefficients of the correct person are in red lines and its training samples are in yellow rectangles. Images are normalized into the same mean and ℓ_2-norm.

The proposed sparse- and dense-hybrid representation (SDR) alleviates this problem. To represent a query image, every training sample only uses its class-specific component to compete against the others (through a sparse minimization) collaboratively with the non-class-specific component of all training samples (through a dense representation). This makes the sparse code of the proposed SDR completely coincide with the specifically assigned class membership. It overcomes a fundamental limitation of SRC that sufficient representative training samples are required for every class. If a class does not have sufficient training samples to represent some variations (non-class-specific component) of its query image, they are represented by the non-class-specific component of other classes and their class memberships do not participate in the classification competition. This is visible in the second row of Fig. 2, which shows an example of the proposed SDR where the top and the fourth significant coefficients are for the training samples of the same identity as the query image.

3 Dictionary Decomposition

The sparse- and dense-hybrid representation (SDR) presented in the last section requires a class-specific dictionary $\mathbf{A}$ and a non-class-specific dictionary $\mathbf{B}$ decomposed from training data $\mathbf{D}$. Towards this end, we apply the proposed conceptual SDR model (9) to every training image in $\mathbf{D} = [\mathbf{d}_1, ... \ \mathbf{d}_k, ... \ \mathbf{d}_n]$ as

$$\mathbf{d}_k = \mathbf{A}\boldsymbol{\alpha}_k + \mathbf{B}\mathbf{x}_k + \mathbf{e}_k, \quad k = 1, \ ..., \ n. \tag{14}$$

Here we set $\boldsymbol{\alpha}_k$ be the sparsest vector as $\boldsymbol{\alpha}_k(j) = 1$ if $j = k$ and $\boldsymbol{\alpha}_k(j) = 0$ if $j \neq k$ for $j, k = 1, 2, ...n$. This is plausible as full information of training sample $\mathbf{d}_k$ can be utilized in the right side of equation (14). Stacking vectors $\mathbf{d}_k$, $\boldsymbol{\alpha}_k$, $\mathbf{x}_k$, $\mathbf{e}_k$ of all different k in (14) into respective matrices yields

$$\mathbf{D} = \mathbf{A} + \mathbf{B}\mathbf{X} + \mathbf{E} \tag{15}$$

The sparse coefficients of $\mathbf{A}$ disappear because $[\boldsymbol{\alpha}_1, \ ..., \ \boldsymbol{\alpha}_n]$ forms an identity matrix.

To realize the proposed SDR framework, we further propose to decompose the training data $\mathbf{D}$ according to the model (15) into a class-specific dictionary $\mathbf{A}$, a non-class-specific dictionary $\mathbf{B}$ and a sparse noise or corruption $\mathbf{E}$, where $\mathbf{X}$ helps the dictionary $\mathbf{B}$ have good representation (prediction) power. If the training data are meaningful, we can reasonably assume $\mathbf{D}$ be a full rank matrix, i.e. $\text{rank}(\mathbf{D}) = \min(m, n)$. This is true with the existence of the random noise in the training data. As the training data $\mathbf{D}$ does not contain all kinds but only a particular type of images – human faces for example, $\mathbf{D} - \mathbf{E}$ should be a low rank matrix. Consequently, $\mathbf{A}$, $\mathbf{BX}$ and $\mathbf{B}$ are all low rank matrices. Therefore, the dictionary decomposition (15) is regularized by

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{X},\mathbf{E}} \text{rank}(\mathbf{A}) + \lambda \text{rank}(\mathbf{B}) + \tau\|\mathbf{X}\|_F^2 + \eta\|\mathbf{E}\|_0 \\ &\text{s.t. } \mathbf{D} = \mathbf{A} + \mathbf{B}\mathbf{X} + \mathbf{E}. \end{aligned} \tag{16}$$

However, the minimization of the ranks of $\mathbf{A}$ and $\mathbf{B}$ and the number of nonzero elements in $\mathbf{E}$ is a highly non-convex optimization problem, which is difficult to solve. Fortunately, it is proven in [41] that the solution of low-rank matrix recovery problem can be well approximated by replacing the rank operator with the nuclear norm $\|\cdot\|_*$ and replacing the ℓ_0-norm with the ℓ_1-norm, which turn it into a convex optimization problem. It is proven that this convex relaxed optimization well recovers the low-rank matrix and the sparse error if the rank of the matrix to be minimized is not too high and the respective error matrix is sparse [41]. A number of algorithms [41], [54], [55] are proposed to solve this convex relaxed problem. Therefore, we relax the regularization of the dictionary decomposition to

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{X},\mathbf{E}} \|\mathbf{A}\|_* + \lambda\|\mathbf{B}\|_* + \tau\|\mathbf{X}\|_F^2 + \eta\|\mathbf{E}\|_1 \\ &\text{s.t. } \mathbf{D} = \mathbf{A} + \mathbf{B}\mathbf{X} + \mathbf{E}. \end{aligned} \tag{17}$$

The squared Frobenius norm $\|\mathbf{X}\|_F^2$ is the sum of the squared ℓ_2-norms, $\|\mathbf{x}_k\|_2^2$. λ, τ and η are parameters to balance the minimization of the four terms.

The decomposition regularization (17) can be solved by minimizing two of the four unknowns with fixed others iteratively. Some initial values are required for the iterative minimization. The random sparse noise is initialized as a null matrix, $\mathbf{E} = \mathbf{0}$. Vector $\mathbf{x}_k$ is a dense representation over the non-class-specific dictionary $\mathbf{B}$. We initialize it as the sparsest vector and let it approach a dense vector during the subsequent learning process. Thus, $\mathbf{x}_k(j) = 1$ if $j = k$ and $\mathbf{x}_k(j) = 0$ if $j \neq k$ for $j, k = 1, 2, ...n$. Thus, $\mathbf{X}$ is initialized as an identity matrix $\mathbf{X} = \mathbf{I}$. Although we decompose the training data $\mathbf{D}$ into four parts $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$ and $\mathbf{E}$, only the two dictionaries $\mathbf{A}$, $\mathbf{B}$ are utilized by the proposed SDR. The above initialization of $\mathbf{E}$ and $\mathbf{X}$ yields $\mathbf{D} = \mathbf{A} + \mathbf{B}$. Thus, the dictionaries $\mathbf{A}$ and $\mathbf{B}$ have not lost any information from the training data $\mathbf{D}$ in the above initialization of $\mathbf{E}$ and $\mathbf{X}$.

As the decomposition regularization (17) involves two low-rank matrices $\mathbf{A}$ and $\mathbf{B}$ simultaneously, there is no reason why $\mathbf{A}$ must capture the class-specific information while $\mathbf{B}$ the non-class-specific one. Therefore, knowledge about the different roles of these four matrices in the proposed SDR must be applied to help decompose the training data $\mathbf{D}$ properly into meaningful $\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$ and $\mathbf{E}$ for the proposed SDR.

Towards this end, we try to allot all possible class-specific information we can have to $\mathbf{A} = [\mathbf{A}_1, ..., \mathbf{A}_i, ..., \mathbf{A}_c]$, $\mathbf{A}_i \in \Re^{m \times n_i}$. If we let $\mathbf{A}_i$ be some kind of class-conditional center such as the mean or median of the training samples of class i, $\mathbf{A}$ should have captured the most significant class-specific information. This is supported by the fact that all approaches of the linear discriminant analysis only utilize the class-conditional mean as the class-specific information. For a better representation power of $\mathbf{A}_i$, this work applies singular value decomposition (SVD) on every class-specific training data $\mathbf{D}_i$

$$\mathbf{D}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T, \tag{18}$$

and lets $\mathbf{A}_i$ be the reconstructed images by the singular vectors corresponding to the largest singular value

$$\mathbf{A}_i = \mathbf{U}_i(1:m, 1)\mathbf{\Sigma}_i(1,1)\mathbf{V}_i(1:n_i, 1)^T. \tag{19}$$

This makes $\mathbf{A}_i$ the lowest rank (rank 1) and the best approximation of $\mathbf{D}_i$ in terms of the Frobenius norm of the difference between $\mathbf{A}_i$ and $\mathbf{D}_i$. Therefore, $\mathbf{A}$ captures the most significant class-specific information with the lowest possible rank, $\text{rank}(\mathbf{A}) = c$. In other words, it allots the least amount of the most significant class-specific information of the training data $\mathbf{D}$ into $\mathbf{A}$. The matrix $\mathbf{B} = \mathbf{D} - \mathbf{A}$ then contains all non-class-specific information and the remaining class-specific information. Obviously, $\text{rank}(\mathbf{B}) = \min(m, n) - c$. In most applications we have in fact $\text{rank}(\mathbf{B}) \gg \text{rank}(\mathbf{A})$. Thus, the subsequent dictionary learning should target at transferring the remaining class-specific information from $\mathbf{B}$ to $\mathbf{A}$.

We first separate the remaining class-specific information from $\mathbf{B}$ by applying the decomposition regularization (17) with fixed $\mathbf{A}$ and $\mathbf{X}$

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{T}} \lambda\|\mathbf{B}\|_* + \delta\|\mathbf{T}\|_1 \\ &\text{s.t. } \mathbf{D} - \mathbf{A} = \mathbf{B}\mathbf{X} + \mathbf{T}. \end{aligned} \tag{20}$$

Symbols $\mathbf{E}$ and η of (17) are changed to $\mathbf{T}$ and δ in (20) to indicate that more component than the random sparse noise should be captured by $\mathbf{T}$ to enable the information transfer. This can be achieved by assigning a smaller weight δ to $\|\mathbf{T}\|_1$.

The minimization of $\text{rank}(\mathbf{B})$ minimizes $\text{rank}(\mathbf{B}\mathbf{X})$ as $\text{rank}(\mathbf{B}\mathbf{X}) \leq \text{rank}(\mathbf{B})$. Thus, the optimization (20) decomposes $\mathbf{D} - \mathbf{A}$ into a sparse matrix $\mathbf{T}$ and a low rank matrix $\mathbf{B}\mathbf{X}$. It coincides with the dense representation part of the proposed SDR (9), which helps the non-class-specific dictionary $\mathbf{B}$ better represent (predict) unknown query image in SDR than directly taking $\mathbf{B}\mathbf{X}$ as the dictionary. Given a small parameter δ, the sparse matrix $\mathbf{T}$ will capture not only the random sparse noise but also the remaining class-specific information in $\mathbf{D} - \mathbf{A}$.

To make the problem (20) solvable, an auxiliary variable $\mathbf{J}$ is introduced to convert (20) to the following equivalent optimization problem:

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{J},\mathbf{T}} \lambda\|\mathbf{J}\|_* + \delta\|\mathbf{T}\|_1 \\ &\text{s.t. } \mathbf{D} - \mathbf{A} = \mathbf{B}\mathbf{X} + \mathbf{T},\ \mathbf{B} = \mathbf{J}. \end{aligned} \tag{21}$$

This can be solved by the ALM method [52] that converts (21) to minimizing an unconstrained function:

$$\begin{aligned} \lambda\|\mathbf{J}\|_* \quad &+ \quad \delta\|\mathbf{T}\|_1 + \text{tr}(\mathbf{Y}_1^T(\mathbf{D} - \mathbf{A} - \mathbf{B}\mathbf{X} - \mathbf{T})) \\ &+ \quad \frac{\mu}{2}\|\mathbf{D} - \mathbf{A} - \mathbf{B}\mathbf{X} - \mathbf{T}\|_F^2 + \text{tr}(\mathbf{Y}_2^T(\mathbf{B} - \mathbf{J})) \\ &+ \quad \frac{\mu}{2}\|\mathbf{B} - \mathbf{J}\|_F^2, \end{aligned} \tag{22}$$

where $\mathbf{Y}_1, \mathbf{Y}_2$ are the Lagrange multipliers, and μ is a penalty parameter. Algorithm 2 summarizes the solution to problem (20) where step 1 is solved via the Singular Value Thresholding [56].

Now by fixing $\mathbf{A}$ and $\mathbf{B}$, we optimize $\mathbf{X}$ and $\mathbf{T}$ by employing the regularization (17) again:

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{T}} \tau\|\mathbf{X}\|_F^2 + \delta\|\mathbf{T}\|_1 \\ &\text{s.t. } \mathbf{D} - \mathbf{A} = \mathbf{B}\mathbf{X} + \mathbf{T}. \end{aligned} \tag{23}$$

Again, the ALM method is applied to covert (23) into an unconstrained minimization function

$$\begin{aligned} \tau\|\mathbf{X}\|_F^2 \quad &+ \quad \delta\|\mathbf{T}\|_1 + \text{tr}(\mathbf{Y}^T(\mathbf{D} - \mathbf{A} - \mathbf{B}\mathbf{X} - \mathbf{T})) \\ &+ \quad \frac{\mu}{2}\|\mathbf{D} - \mathbf{A} - \mathbf{B}\mathbf{X} - \mathbf{T}\|_F^2 \end{aligned} \tag{24}$$

where $\mathbf{Y}$ is the Lagrange multiplier and μ a penalty parameter. The inexact ALM method to solve (24) is outlined in Algorithm 3.

With the optimized $\mathbf{B}$ and $\mathbf{X}$, the remaining class-specific information in $\mathbf{D} - \mathbf{A}$ is separated from $\mathbf{B}\mathbf{X}$. The last step of the proposed dictionary decomposition is to transfer it to $\mathbf{A}$. This is achieved by using the regularization (17) again with fixed $\mathbf{B}$ and $\mathbf{X}$

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \eta\|\mathbf{E}\|_1 \\ &\text{s.t. } \mathbf{D} - \mathbf{B}\mathbf{X} = \mathbf{A} + \mathbf{E}. \end{aligned} \tag{25}$$

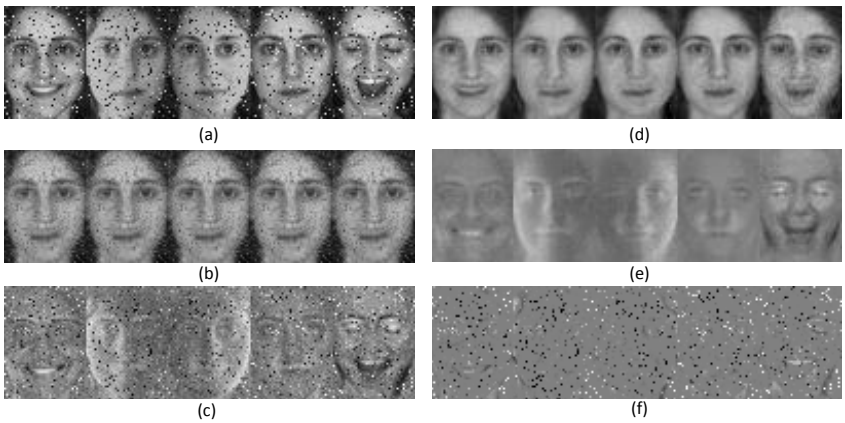


Fig. 3. Sample images produced by the proposed SLR: images of a person in $\mathbf{D}$ corrupted by random noise (a); their initial components assigned to $\mathbf{A}$ (b) and $\mathbf{B}$ (c); final images in $\mathbf{A}$ (d), $\mathbf{BX}$ (e) and $\mathbf{E}$ (f).

Algorithm 2: Solving Problem(20) by Inexact ALM

Input: $\mathbf{D}$, $\mathbf{A}$, $\mathbf{X}$ and parameter λ and δ.
Initialize: $\mathbf{B} = \mathbf{J} = 0$, $\mathbf{T} = 0$, $\mathbf{Y_1} = 0$, $\mathbf{Y_2} = 0$, $\mu = 10^{-3}$, $\mu_{max} = 10^6$, $\rho = 1.5$, and $\epsilon = 10^{-6}$.
while *not converged* **do**

1. fix the others and update $\mathbf{J}$ by
$$\mathbf{J} = \arg\min \tfrac{\lambda}{\mu}\|\mathbf{J}\|_* + \tfrac{1}{2}\|\mathbf{J} - (\mathbf{B} + \mathbf{Y_2}/\mu)\|_F^2$$
2. fix the others and update $\mathbf{B}$ by
$$\mathbf{B} = ((\mathbf{D}-\mathbf{A}-\mathbf{T})\mathbf{X}^T + \mathbf{J} + (\mathbf{Y_1}\mathbf{X}^T - \mathbf{Y_2})/\mu)(\mathbf{X}\mathbf{X}^T + \mathbf{I})^{-1}$$
3. fix the others and update $\mathbf{T}$ by
$$\mathbf{T} = \arg\min \tfrac{\delta}{\mu}\|\mathbf{T}\|_1 + \tfrac{1}{2}\|\mathbf{T} - (\mathbf{D} - \mathbf{A} - \mathbf{BX} + \mathbf{Y_1}/\mu)\|_F^2$$
4. update the multipliers
$$\begin{aligned}\mathbf{Y}_1 &= \mathbf{Y}_1 + \mu(\mathbf{D} - \mathbf{A} - \mathbf{BX} - \mathbf{T})\\ \mathbf{Y}_2 &= \mathbf{Y}_2 + \mu(\mathbf{B} - \mathbf{J})\end{aligned}$$
5. update μ by $\mu = min(\rho\mu, \mu_{max})$
6. check the convergence conditions:
$$\|\mathbf{D} - \mathbf{A} - \mathbf{BX} - \mathbf{T}\|_F^2 < \epsilon \text{ and } \|\mathbf{B} - \mathbf{J}\|_F^2 < \epsilon$$

end
Output: $\mathbf{B}$

Algorithm 3: Solving Problem(23) by Inexact ALM

Input: $\mathbf{D}$, $\mathbf{A}$, $\mathbf{B}$, and parameter τ and δ.
Initialize: $\mathbf{T} = 0$, $\mathbf{Y} = 0$, $\mu = 10^{-3}$, $\mu_{max} = 10^6$, $\rho = 1.5$, and $\epsilon = 10^{-6}$.
while *not converged* **do**

1. fix the others and update $\mathbf{X}$ by
$$\mathbf{X} = \mu(2\tau\mathbf{I} + \mu\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{D} - \mathbf{A} - \mathbf{T} + \mathbf{Y}/\mu)$$
2. fix the others and update $\mathbf{T}$ by
$$\mathbf{T} = \arg\min \tfrac{\delta}{\mu}\|\mathbf{T}\|_1 + \tfrac{1}{2}\|\mathbf{T} - (\mathbf{D} - \mathbf{A} - \mathbf{BX} + \mathbf{Y}/\mu)\|_F^2$$
3. update the multiplier
$$\mathbf{Y} = \mathbf{Y} + \mu(\mathbf{D} - \mathbf{A} - \mathbf{BX} - \mathbf{T})$$
4. update μ by $\mu = min(\rho\mu, \mu_{max})$
5. check the convergence conditions:
$$\|\mathbf{D} - \mathbf{A} - \mathbf{BX} - \mathbf{T}\|_F^2 < \epsilon$$

end
Output: $\mathbf{X}$

As (25) is a standard low-rank matrix recovery problem, it can be solved by the algorithm in [41]. To ensure only random sparse noise being captured by $\mathbf{E}$, a high weight η should be assigned to $\|\mathbf{E}\|_1$.

The optimizations (20), (23) and (25) need iterate a few times due to two unknowns in the form of $\mathbf{BX}$. Experiments show that only 3 to 5 iterations yield the stable best results. Thus, all experiments of this work iterate (20), (23) and (25) 4 times. We call the above proposed procedure the supervised low-rank (SLR) dictionary decomposition, which is summarized in Algorithm 4. It utilizes the low-rank matrix recovery to transfer information from the supervised assigned dictionary $\mathbf{B}$ to $\mathbf{A}$. The matrix $\mathbf{T}$ serves as a medium of the information transfer, which leads the final dictionary $\mathbf{A}$ containing more class-specific information than that used in the conventional linear discriminant analysis that is equivalent to the initial $\mathbf{A}$. The matrix $\mathbf{X}$ improves the quality of dictionary $\mathbf{B}$ for the representation (prediction) of unknowns in the proposed SDR framework. The proposed SLR dictionary decomposition not only delivers a class-specific dictionary $\mathbf{A}$ and a non-class-specific dictionary $\mathbf{B}$, it also alleviates the other fundamental problem of SRC – corrupted training data, since the random sparse noise of training data $\mathbf{E}$ is removed from the two learnt dictionaries.

Fig. 3 visualizes some results of the proposed SLR dictionary decomposition on AR database. Training images of one of 100 persons corrupted by strong sparse noise in $\mathbf{D}$ are shown in Fig. 3 (a). Their initial assignments to $\mathbf{A}$ and $\mathbf{B}$ are shown in Fig. 3 (b) and (c), respectively. The final components of Fig. 3 (a) in $\mathbf{D}$ decomposed by the proposed SLR in $\mathbf{A}$, $\mathbf{BX}$ and $\mathbf{E}$ are shown in Fig. 3 (d), (e) and (f), respectively. Fig. 3 also serves as a visual comparison of the dictionaries decomposed by the proposed SLR to those used in SSRC [47], a further development from ESRC [46]. SSRC uses the class-wise means of $\mathbf{D}$ as one

Algorithm 4: SLR Dictionary Decomposition

Input: $\mathbf{D} = [\mathbf{D}_1, \ ..., \ \mathbf{D}_i, \ ..., \ \mathbf{D}_c]$ and parameter κ.
Initialize: $\mathbf{X} = \mathbf{I}$, $\mathbf{D}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^T$,
$\mathbf{A}_i = \mathbf{U}_i(1:m,1)\mathbf{\Sigma}_i(1,1)\mathbf{V}_i(1:n_i,1)^T$,
$\mathbf{A} = [\mathbf{A}_1, \ ..., \ \mathbf{A}_i, \ ..., \ \mathbf{A}_c]$.
for *1:*κ **do**
1. obtain $\mathbf{B}$ by solving the optimization problem (20) using Algorithm 2.
2. update $\mathbf{X}$ by solving the optimization problem (23) using Algorithm 3.
3. update $\mathbf{A}$ by solving standard low-rank matrix recovery problem (25).
end
Output: $\mathbf{A}$ and $\mathbf{B}$

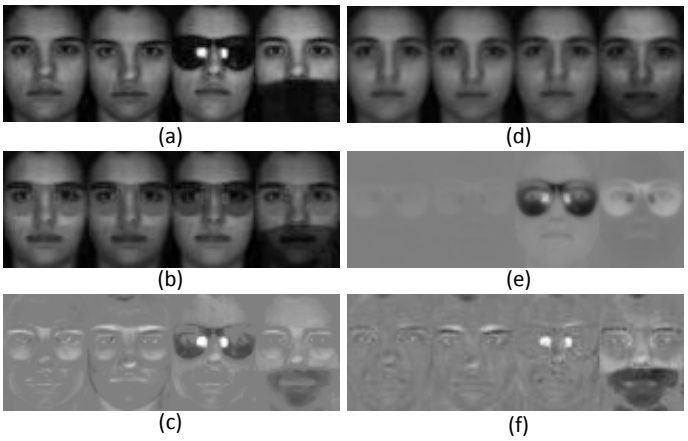


Fig. 4. Examples of RPCA and the proposed SLR: training images (a); low-rank (b) and sparse (c) images from RPCA; class-specific (d), non-class-specific (e) and sparse (f) images decomposed by SLR.

dictionary and the class-wise centralized $\mathbf{D}$ as the other. They are similar to Fig. 3 (b) and (c) except for a scaling factor.

Fig. 4 shows a further visual example of the proposed SLR comparing to RPCA. The data set contains 301 face images from AR database, where 1 wears scarf and each of 100 persons has 1 image wearing sunglasses and 2 undisguised images. Fig. 4 (b) and (c) show that sunglasses and scarf in Fig. 4 (a) are not fully removed from the dictionary but shared by it and the sparse noise by RPCA. In contrast, they are fully removed from the class-specific dictionary by SLR as shown in Fig. 4 (d). Moreover, sunglasses, though much small than scarf, is allotted in the non-class-specific dictionary. This is desirable as over 30% samples in the data set wear sunglasses and so is expected for the query images. Allotting it in $\mathbf{B}$ increases its representation power. This improves rather than damages the classification.

4 Experiments

The proposed SDR framework with the proposed SLR dictionary decomposition algorithm, SDR-SLR, is evaluated on 4 face databases: CMU Multi-PIE [57], Extended Yale B [58], AR [59] and FERET [60]. It is compared with related state-of-the-art SRC [21], ESRC [46], SSRC [47], LR (Low-Rank SRC) and LRSI (LR with structural incoherence) [42] as well as baseline classifiers NN (nearest neighbor) and LRC [17]. The parameters of the proposed SDR-SLR approach are fixed over all experiments of this paper to $\beta = \gamma = 10$ for the SDR and $\kappa = 4$, $\lambda = 1$, $\tau = 0.01$, $\delta = 0.8\nu$ and $\eta = 1.2\nu$ for SLR, where $\nu = 1/\sqrt{max(size(\mathbf{D}))}$ [54], though better performance may be achieved if they are fine tuned to fit each specific experiment.

The CMU Multi-PIE database contains face images captured in 4 sessions with variations in illumination, expression and pose. The first 105 subjects that appear in all 4 sessions are used in the experiments. Images are cropped based on the eye locations provided by [61] and down-sampled to 50×40 pixels.

The cropped Extended Yale B database has 2,414 frontal-face images of size 192×168 from 38 subjects, captured under different lighting conditions. 64 images per subject are divided into 5 subsets according to the angle of the light direction with 7, 12, 12, 14 and 19 images respectively in subset 1 to 5.

The AR database used in this paper has 2600 frontal-face images of size 165×120 from 100 subjects (50 males and 50 females). In each of two separate sessions, 7 undisguised images with expression or illumination variation, 3 images in sunglasses and 3 images in scarf disguise are taken from each subject.

A subset of FERET database contains 1024 images from 256 subjects with 4 frontal images per subject. Images are normalized by an affine transformation, scaled and cropped to the size of 121×121. Fig. 5 shows the normalized and cropped images.

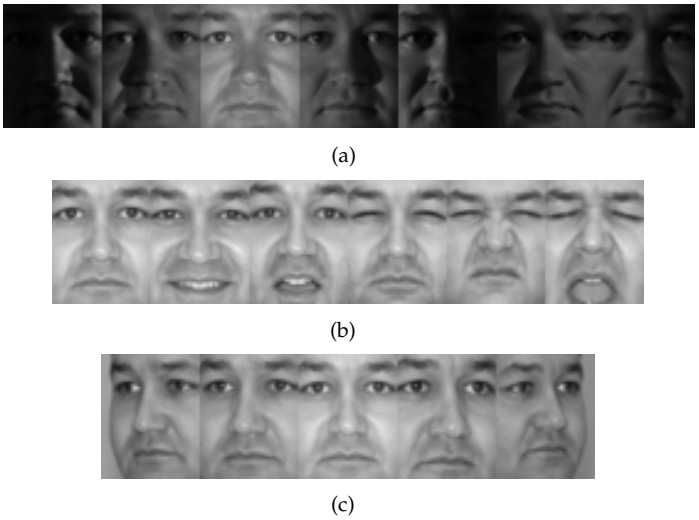
Fig. 5. Normalized images from FERET database.

4.1 Face Recognition on Single Variation

This subsection tests the effectiveness of the proposed SDR-SLR on isolated single type of variation: illumination, expression or pose. Misalignment is not tested as it can be alleviated by using alignment insensitive features such as histogram of gradient [62] or LBP [63], [64]. Recent efforts that recover the misalignment can be found in [27], [65].

Experiment 1(illumination): following the procedure in [57], 18 flash-only images are generated as the difference between flash images (illuminations {1-18}) and non-flash image (illumination {0}) of the Multi-PIE database as shown in Fig. 6(a). For each of these 18 different illuminations, 4 frontal images with neutral expression per subject are chosen from 4 different sessions, which produces $18 \times 4 \times 105 = 7560$ images for training and testing. For each subject, t different illuminations are randomly selected for training and the rest $18-t$ illuminations are used for testing. The averaged recognition rates of

10 runs are plotted against the reducing t in Fig. 7. Much to our surprise, ESRC, SSRC, LR and LRSI, which try to solve the problems of SRC, in fact underperform SRC. Nevertheless, the proposed SDR-SLR consistently outperforms the others in all cases where the performance gain increases with decreasing illumination variations in the training data.



(a) (b) (c)

Fig. 6. Sample images in Multi-PIE database with variations of (a) illumination, (b) expression, (c) pose.

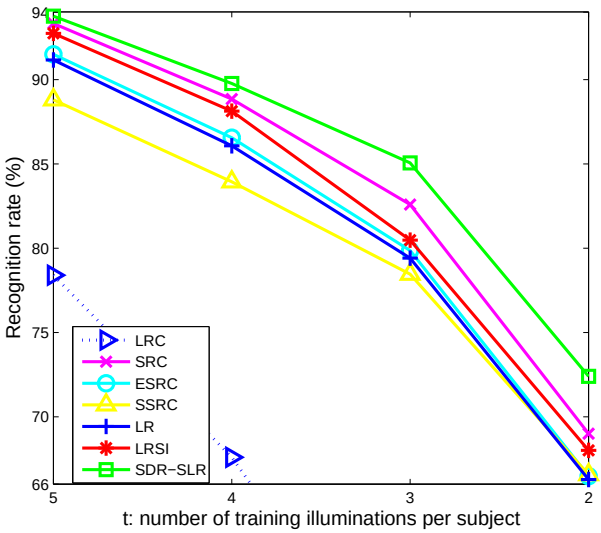


Fig. 7. Face recognition rate versus number of training illuminations per subject on Multi-PIE database.

Experiment 2(illumination): to further verify the performances of various methods in tackling the illumination variation, another experiment is done on Extended Yale B database. For each subject, 2 subsets are randomly chosen for training and the other 3 subsets for testing. Different image sizes are tested. The average recognition rates of 10 runs in Table 1 further show that ESRC, SSRC, LR and LRSI underperform SRC on data set that has only illumination variation. Again, the proposed SDR-SLR visibly outperforms the others consistently for all image sizes.

TABLE 1
Face Recognition Rate on Extended Yale B

dimensions	810	1400	2016
NN	42.1%	43.9%	44.6%
LRC [17]	72.8%	73.1%	73.2%
SRC [21]	77.9%	78.5%	78.7%
ESRC [46]	77.6%	78.5%	78.6%
SSRC [47]	74.8%	75.2%	76.5%
LR [42]	77.3%	77.6%	78.0%
LRSI [42]	77.8%	78.0%	78.5%
SDR-SLR	**81.7**%	**84.3**%	**84.5**%

Experiment 3(expression): all frontal images with illumination {7} are taken from the Multi-PIE database. The 6 expressions (neutral, smile, surprise, squint, disgust and scream) are shown in Fig. 6(b). For each subject, t different expressions are randomly selected for training and the rest $6-t$ expressions are used for testing. The averaged recognition rates of 10 runs are plotted against the reducing t in Fig. 8. It shows that LR and LRSI performs similar or even inferior to SRC as the database is uncorrupted. As no similar expression of the query image appears in the training samples of the same subject, ESRC, SSRC and SDR-SLR significantly outperform SRC. The proposed SDR-SLR consistently and visibly performs the best for all values of t where the performance gain increases with decreasing expression variations in the training data.

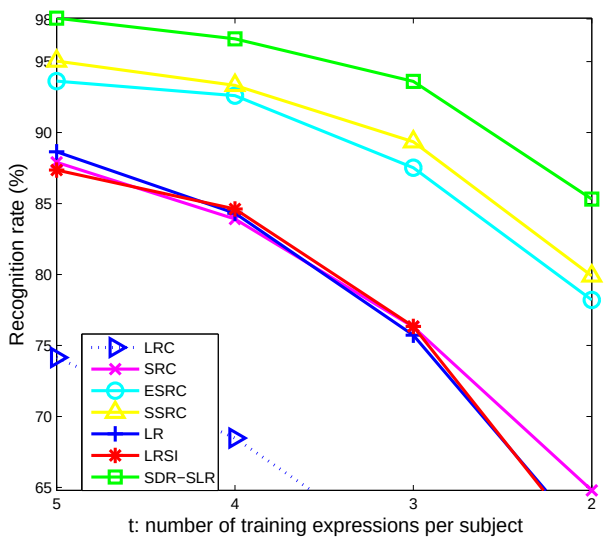


Fig. 8. Face recognition rate versus number of training expressions per subject on Multi-PIE database.

Experiment 4(pose): 20 images of neutral expression from 4 sessions captured by 5 cameras (from -30% to +30%) and flashed by the capturing camera are taken from Multi-PIE database as shown in Fig. 6(c). For each subject, t different poses are randomly chosen for training and the rest $5-t$ poses are used for testing. Table 2 details the average results over 10 runs. LR and LRSI only slightly outperform SRC on the uncorrupted data. Similar to the experiment 3, ESRC, SSRC and SDR-SLR

TABLE 2
Face Recognition Rate on the Multi-PIE Face Database with Pose Variation.

t	4	3	2
NN	17.0%	13.7%	11.4%
LRC [17]	41.9%	29.5%	18.4%
SRC [21]	65.8%	57.8%	45.2%
ESRC [46]	73.0%	69.0%	57.3%
SSRC [47]	79.3%	73.8%	61.2%
LR [42]	66.2%	59.4%	46.6%
LRSI [42]	66.6%	58.6%	45.2%
SDR-SLR	**87.2**%	**81.4**%	**67.1**%

TABLE 4
Recognition Rate of Fewer Training Samples of AR Dataset

train. samples	6	5	4	3
NN	65.4%	60.4%	53.5%	46.1%
LRC [17]	72.8%	67.4%	60.5%	53.7%
SRC [21]	91.7%	88.7%	87.7%	82.3%
ESRC [46]	90.4%	89.4%	88.4%	85.1%
SSRC [47]	90.3%	89.4%	89.0%	84.7%
LR [42]	91.9%	87.8%	86.7%	79.5%
LRSI [42]	91.7%	88.6%	87.3%	79.7%
SDR-SLR	**96.4**%	**96.3**%	**95.6**%	**92.1**%

significantly outperform SRC, where the proposed SDR-SLR consistently and visibly performs the best for all values of t.

4.2 Face Recognition on Mixed Variations

This subsection tests the effectiveness of the proposed SDR-SLR on databases whose face images have mixed types of variations without image corruption.

Experiment 5: all undisguised images from Session 1 of the AR database are used for training and those from Session 2 are used for testing. Images down-sampled to 4 different sizes are tested. The results are recorded in Table 3. Unfortunately, ESRC, SSRC, LR and LRSI again do not outperform SRC though the data have the both expression and illumination variations. Probably the training and testing sets, though taken from two separate sessions, have very similar variations. Nevertheless, the proposed SDR-SLR approach again visibly outperforms the others consistently for all image sizes.

TABLE 3
Recognition Rate on All Undisguised Images of AR Dataset

dimensions	540	850	1200	2200
NN	69.5%	70.4%	71.2%	71.8%
LRC [17]	74.1%	75.2%	76.0%	76.4%
SRC [21]	90.7%	91.6%	92.4%	92.8%
ESRC [46]	90.9%	90.8%	91.4%	91.8%
SSRC [47]	87.7%	90.7%	90.7%	90.7%
LR [42]	90.8%	91.3%	91.7%	92.8%
LRSI [42]	91.1%	91.3%	91.6%	92.5%
SDR-SLR	**96.4**%	**96.7**%	**96.9**%	**97.6**%

Experiment 6: to show the cases where the problems of SRC become severe, we repeat the above experiment for the image dimensions of 2200 with reduced number of training samples per subject to 6, 5, 4, and 3. The average results of 10 runs in Table 4 show that, except for the case of 6 training samples, LR and LRSI underperform SRC as the training data have no corruption. ESRC and SSRC start outperforming SRC from 5 training samples, yet marginally. The best accuracy gain of ESRC or SSRC over SRC is 2.8% at 3 training samples. In contrast, that of the proposed SDR-SLR reaches almost 10%.

Experiment 7: for each subject of the FERET data set, we randomly pick out two images, one for training and the other for testing. Then, the half of the remaining 512 images is randomly taken for training and the rest are for testing. Thus, there are 512 images in training and the other 512 images in testing with at least one in training and another in testing per subject. Images down-sampled to 4 different sizes are tested. The average results of 10 runs in Table 5 show that LR and LRSI have about the same accuracy as SRC. For this small number of training samples per class, the both ESRC and SSRC outperform SRC with the gains between 2% and 4.1%. Much more significant gains in recognition performance over SRC are achieved by the proposed SDR-SLR consistently over all different image dimensions. They are between 9.2% and 10.7%.

TABLE 5
Face Recognition Rate on FERET Database

dimensions	400	900	1600	2500
NN	68.5%	67.6%	67.0%	66.7%
LRC [17]	68.0%	67.5%	66.9%	66.5%
SRC [21]	80.8%	80.1%	79.9%	79.1%
ESRC [46]	83.5%	83.5%	82.9%	82.4%
SSRC [47]	83.3%	84.2%	81.9%	82.9%
LR [42]	80.9%	79.6%	79.2%	79.0%
LRSI [42]	81.1%	80.0%	79.7%	79.4%
SDR-SLR	**90.0**%	**90.1**%	**90.6**%	**89.5**%

4.3 Face Recognition on Corrupted Data

This subsection tests the effectiveness of various approaches on training data with different types and levels of corruption.

Experiment 8: the undisguised subset of AR database is used. 5 images per subject from Session 1 are randomly chosen as training set and all images from Session 2 are used as test set. All images are resized to 50×44 pixels. For each training image, a certain percentage of its pixels are randomly replaced by noise uniformly distributed between the minimal and the maximal pixel value. The average recognition rate over 10 runs is plotted against the noise level in Fig. 9. It shows that ESRC and SSRC perform about the same as SRC. LR and LRSI slightly outperform SRC thanks to their low-rank recovery of the training data. Fig. 9 demonstrates that the proposed SDR-SLR significantly outperforms all other algorithms consistently for all levels of corruption.

Experiment 9: besides the random pixel corruption, we further test different approaches in coping with random block occlusion on Extended Yale B database. Same as [21], Subsets 1 and 2 containing 719 images are taken for training and Subset 3 containing 455 images is used for testing. Images are resized to 48×42. Also same as [21], different portions of images, from 20% to 50%, are

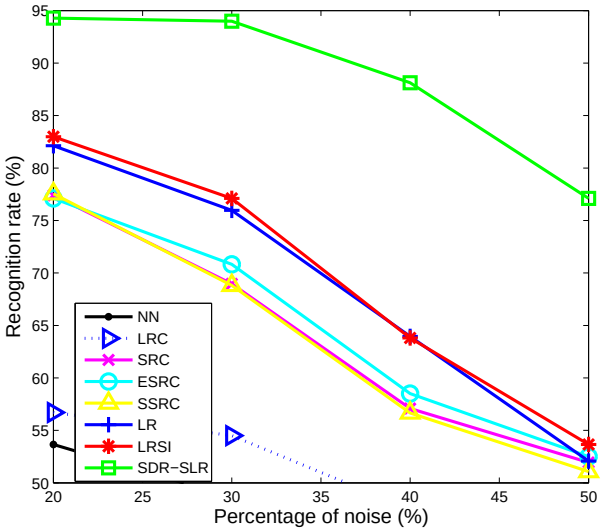


Fig. 9. Face recognition rate versus percentage of uniform noise on AR database.

Fig. 10. Examples of images of Extended Yale B database with different levels of block occlusion. From left to right, 20%, 30%, 40% and 50% of images are occluded, respectively.

occluded by an unrelated image at random locations as shown in Fig. 10. But different from [21] that corrupts all test images, we corrupt half of the training and half of the testing images. The result over 10 runs is plotted against the level of occlusion in Fig. 11.

It seems a surprise that for this corrupted data, LR and LRSI do not consistently have better performance than SRC but ESRC and SSRC do. Possible reasons could be that the same image is used to occlude all 50% training and testing images so that the block occlusion causes many training samples having a same occluded area and even same as many testing images. As a result, the occlusion appears more like a common variation than the random corruption. Nevertheless, the proposed SDR-SLR outperforms all other algorithms consistently for all levels of occlusion.

4.4 Face Recognition with Real Disguised Images

This subsection tests the effectiveness of various approaches in dealing with real possible malicious occlusions in both training and testing samples. Using all images of size 55×40 in AR database,

Experiment 10 considers the following 4 scenarios:

Sunglasses only: for each of the 100 subjects, all 7 undisguised images and 1 image with sunglasses (random chosen) from Session 1 are selected as training set. All remaining undisguised images and images with sunglasses are used for testing. So, there are 8 training and 12 testing images per subject.

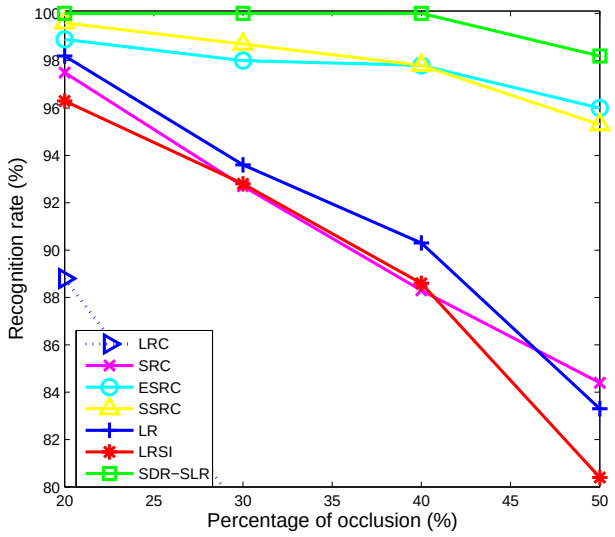


Fig. 11. Face recognition rate versus percentage of occlusion on extended Yale B database.

Scarf only: replace images with sunglasses in the above scenario by images with scarf.

Mixed 1 (Sunglasses and Scarf): for each subject, all undisguised images, 1 image with sunglasses (random chosen) and 1 image with scarf (random chosen) from Session 1 are selected as training set. All remaining images are used for testing. So, there are 9 training and 17 testing images per subject.

Mixed 2 (Sunglasses or Scarf): 1 image randomly taken from the 6 disguised images per subject and all undisguised images from Session 1 are used for training. All remaining images are used for testing. So, there are 8 training and 18 testing images per subject.

The first 3 scenarios are exactly the same as those used in [42], [47]. The forth scenario is new and more challenge because every class has two different disguise types in its testing data, one of which, though presents in some other classes, is absent in its training data. We repeat each scenario three times and the average results are recorded in Table 6. It shows that LR or LRSI is consistently but very marginally better than SRC. The both ESRC and SSRC outperform SRC more visibly, where SSRC is consistently the second best performer, which gains accuracy over SRC between 1.1% and 6.6%. The proposed SDR-SLR consistently performs the best and its accuracy gain over SRC ranges from 7.5% to 13.1%.

5 Conclusion

Sparse representation shows some merits in holistic image classification. In seeking a sparse representation of a query image, every sample competes against the others to gain its share. This well matches the general classification objective that only one class should stand out from the rest. However, all information in images participates in

TABLE 6
Face Recognition Rate on the AR Database.

Scenario	Sunglass	Scarf	Mixed 1	Mixed 2
NN	51.4%	49.0%	41.8%	35.8%
LRC [17]	69.8%	64.8%	64.2%	47.3%
SRC [21]	88.8%	86.6%	85.8%	79.9%
ESRC [46]	89.6%	89.5%	88.9%	84.5%
SSRC [47]	89.9%	90.1%	89.3%	86.5%
LR [42]	88.6%	86.8%	86.1%	79.9%
LRSI [42]	89.0%	86.7%	86.3%	80.2%
SDR-SLR	**96.3**%	**94.5**%	**95.3**%	**93.0**%

this competition. There is no reason why the significant coefficients must coincide with a specific one of many different forms of class grouping, especially for some applications where the class-specific information takes up only a very small portion of that in images. As a result, SRC requires sufficient representative samples for every class, with which the non-class-specific information can be nulled out. This requirement may not be fulfilled in many computer vision problems such as face recognition. Another problem of SRC is the corrupted training data though it is robust to the corrupted query image thanks to the ℓ_1 minimization of the error.

This work alleviates the two fundamental problems of SRC by a sparse- and dense-hybrid representation (SDR) based on a supervised low-rank (SLR) dictionary decomposition/learning. In the proposed SDR framework, every sample only uses its class-specific component to compete against the others collaboratively with the non-class-specific component of all samples. This makes the sparse code of the proposed SDR completely coincide with the specifically assigned class membership. The co-representation by the non-class-specific component largely relaxes the requirement of representative samples for every class. The class-specific dictionary decomposed by the proposed SLR captures more information than that used in the linear discriminant analysis. The sparse outlier pixels and occlusions of the training data are also separated from the two decomposed low-rank dictionaries. This alleviates the second problem of SRC with corrupted training data.

Extensive experiments on 4 face image databases have verified the effectiveness and advancement of the proposed SDR-SLR approach. It consistently and visibly outperforms SRC and its related extensions in all experiments, whether the training data are corrupted or not and whether every class has sufficient representative training samples or not. In case the training data are corrupted or are not representative for some classes, the performance gains of the proposed SDR-SLR approach are significant.

REFERENCES

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

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

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711 –720, July 1997.

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

[5] B. Moghaddam, "Principal manifolds and probabilistic subspace for visual recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 6, pp. 780–788, June 2002.

[6] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *Electronics Letters*, vol. 42, no. 19, pp. 1089–1090, Sep. 2006.

[7] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 3, pp. 328 –340, Mar. 2005.

[8] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[9] X. D. Jiang, B. Mandal, and A. C. Kot, "Complete discriminant evaluation and feature extraction in kernel space for face recognition," *Machine Vision and Applications*, vol. 20, no. 1, pp. 35–46, Jan. 2009.

[10] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "Kpca plus lda: A complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[11] X. D. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, March 2011.

[12] S. Z. Li and J. W. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. Neural Networks*, vol. 10, no. 2, pp. 439 –443, Mar. 1999.

[13] J. T. Chien and C. C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 12, pp. 1644 – 1649, Dec. 2002.

[14] J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 11–18.

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

[16] S. Z. Li and S. O. C. Ieee Comp, "Face recognition based on nearest linear combinations," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998, pp. 839–844.

[17] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.

[18] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[19] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[20] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Dec. 2006.

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

[22] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *Proc. European Conf. Computer Vision*, vol. 6316, 2010, pp. 448–461.

[23] Y. N. Liu, F. Wu, Z. H. Zhang, Y. T. Zhuang, and S. C. Yan, "Sparse representation using nonnegative curds and whey," in *IEEE Conf. Computer Vision and Pattern Recognition*, june 2010, pp. 3578–3585.

[24] J. J. Wang, J. C. Yang, K. Yu, F. J. Lv, T. Huang, and Y. H. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367.

[25] R. He, W. S. Zheng, and B. G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[26] J. Lai and X. D. Jiang, "Modular weighted global sparse representation for robust face recognition," *IEEE Signal Processing Letter*, vol. 19, no. 9, pp. 571–574, Sep. 2012.

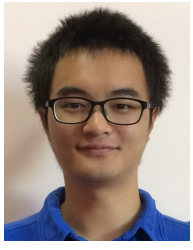
[27] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Hossein, and Y. Ma, "Towards a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[28] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[29] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[30] J. J. Wang, J. C. Yang, K. Yu, F. J. Lv, T. Huang, and Y. H. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367.
[31] Q. Shi, A. Eriksson, A. v. d. Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.
[32] L. Zhang, M. Yang, and X. C. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int'l. Conf. Computer Vision*, 2011, pp. 471–478.
[33] R. Rigamonti, M. A. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1545–1552.
[34] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011.
[35] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.
[36] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 3501–3508.
[37] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.
[38] ——, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems*, 2008.
[39] Z. L. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1697–1704.
[40] M. Yang, D. Zhang, X. C. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int'l Conf. Computer Vision*, 2011, pp. 543–550.
[41] E. J. Candès, X. D. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, Jun. 2011.
[42] C. F. Chen, C. P. Wei, and Y. C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2618–2625.
[43] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2586–2593.
[44] Y. Zhang, Z. Jiang, and L. Davis, "Learning structured low-rank representations for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.
[45] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 1, pp. 171–184, 2013.
[46] W. H. Deng, J. N. Hu, and J. Guo, "Extended src: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 9, pp. 1864–1870, 2012.
[47] ——, "In defense of sparsity based face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.
[48] R. Barsi and D. Jacobs, "Lambertian reflection and linear subspaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 2, pp. 218–233, 2003.
[49] D. Donoho and Y. Tsaig, "Fast solution of l1-norm minimization problems when the solution may be sparse," *IEEE Trans. Information Theory*, vol. 54, no. 11, pp. 4789–4812, 2008.
[50] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.
[51] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, 2008.
[52] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1996.
[53] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.
[54] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *UIUC Technical Report UILU-ENG-09-2215*, 2009.
[55] Z. C. Lin, A. Ganesh, J. Wright, L. Q. Wu, M. M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," in *In Intl. Workshop on Comp. Adv. in Multi-Sensor Adapt. Processing*, 2009.
[56] J. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.
[57] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807 – 813, 2010.
[58] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.
[59] A. Martinez and R. Benavente, "The ar face database," *CVC Technical Report*, no. 24, June 1998.
[60] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 10, pp. 1090–1104, 2000.
[61] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, "A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1788–1794, Jul. 2013.
[62] A. Satpathy, X. D. Jiang, and H. Eng, "Human detection by quadratic classification on subspace of extended histogram of gradients," *IEEE Trans. Image Processing*, vol. 23, no. 1, pp. 287–297, Jan 2014.
[63] J. Ren, X. D. Jiang, and J. Yuan, "Noise-resistant local binary pattern with an embedded error-correction mechanism," *IEEE Trans. Image Processing*, vol. 22, no. 10, pp. 4049–4060, Oct 2013.
[64] A. Satpathy, X. D. Jiang, and H. Eng, "Lbp based edge-texture features for object recognition," *IEEE Trans. Image Processing*, vol. 23, no. 5, pp. 1953–1964, May 2014.
[65] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233 – 2246, 2012.

Xudong Jiang (M'02-SM'06) received the B.Eng. and M.Eng. degrees in Electrical Engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, and the Ph.D. degree in Electrical Engineering from Helmut Schmidt University (HSU), Hamburg, Germany, in 1983, 1986, and 1997, respectively. From 1986 to 1993, he was a Lecturer at UESTC, where he received two Science and Technology Awards from the Ministry for Electronic Industry of China. From 1993 to 1997, he was a Scientific Assistant with HSU. From 1998 to 2004, he was with the Institute for Infocomm Research, A*STAR, Singapore, as a Lead Scientist and the Head of the Biometrics Laboratory, where he developed a system that achieved the most efficiency and the second most accuracy at the International Fingerprint Verification Competition in 2000. He was with the Nanyang Technological University (NTU), Singapore, as a Faculty Member, in 2004, and served as the Director with the Centre for Information Security from 2005 to 2011. He is currently a Tenured Associate Professor with the School of Electrical and Electronic Engineering, NTU. He has authored more than 100 papers, with 19 papers in IEEE Journals: TPAMI (5), TIP (5), TSP (3), SPL(2), SPM, TIFS, TCSVT and TCS-II. He holds seven patents. His current research interests include pattern recognition, computer vision, machine learning, signal/image processing, and biometrics.



Jian Lai received the B.Eng. degree from Zhejiang University, Hangzhou, China in 2009. He is currently pursuing his Ph.D. degree in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include pattern recognition, machine learning and computer vision, with specific interest in face recognition. He is a student member of IEEE.