Classwise Sparse and Collaborative Patch Representation for Face Recognition

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

***Abstract*— Sparse representation has shown its merits in solving some classification problems and delivered some impressive results in face recognition. However, the unsupervised optimization of the sparse representation may result in undesired classification outcome if the variations of the data population are not well represented by the training samples. In this paper, a method of class-wise sparse representation (CSR) is proposed to tackle the problems of the conventional sample-wise sparse representation and applied to face recognition. It seeks an optimum representation of the query image by minimizing the class-wise sparsity of the training data. To tackle the problem of the uncontrolled training data, this paper further proposes a collaborative patch (CP) framework, together with the proposed CSR, named CSR-CP. Different from the conventional patch-based methods that optimize each patch representation separately, the CSR-CP approach optimizes all patches together to seek a CP groupwise sparse representation by putting all patches of an image into a group. It alleviates the problem of losing discriminative information in the training data caused by the partition of the image into patches. Extensive experiments on several benchmark face databases demonstrate that the proposed CSR-CP significantly outperforms the sparse representation-related holistic and patch-based approaches.**

***Index Terms*— Sparse representation, class-wise sparsity, classification, holistic, patch based, face recognition.**

I. Introduction

THE RAPID growing popularity of social network, such as Facebook, Instagram and Pinterest, produces increasing amount of image content shared online. How to organize these data and retrieve information from them attracts many researchers in image processing to the problems of image classification and image understanding. Human faces are probably the most popular image content. Moreover, the rich off-the-shelf face databases provide excellent benchmark data for evaluation of various techniques of image classification and image understanding. Especially, they provide a great test bed for the problems of large number of classes with few training image per classes, which is a common scenario in many real applications.

Holistic approach directly applies the dimensionality reduction and classification techniques to the whole image represented by dense features such as pixel level, HoG [1] or LBP [2], [3]. For dimensionality reduction, many methods are developed based on various criteria. While principal component analysis [4], [5] minimizes the reconstruction error in the low dimensional space, linear discriminative analysis [6], [7] maximizes the discriminant power in the subspace. Locality preserving projection [8] maintains the local relation in the embedding. In [9], the principle and rationale behind the positive role of dimensionality reduction for classification is revealed. Although there are many different classifiers available [10], the suitable ones for high dimensional data are limited. Nearest neighbor (NN) classifier is widely applied for its simplicity. Nearest feature line [11] (NFL) extends the isolated feature points to lines by linking two training samples. Sharing a similar idea, nearest feature plane (NFP) [12], nearest feature space (NFS) [12]–[14] and linear regression classifier (LRC) [15] further expand the training samples to feature planes and subspaces. The commonalities of all these classifiers are that they evaluate the relation between the query image and the training samples of each individual subject separately.

Different from these classifiers, Wright *et al.* [16] proposed a sparse representation based classification (SRC) scheme, which considers the query image as a linear combination of all training samples. It is in general assumed that samples of a specific subject lie in a linear subspace [6], [17]. With this assumption, the query image is expected to be well represented by the training samples of the same subject, which may lead to a sparse representation over all training data. SRC achieves some impressive face recognition performances with sufficient and uncorrupted training data. Inspired by SRC, many extensional works are proposed, such as SRC with nonnegative constraint [18], Gabor feature based SRC [19], Gaussian kernel error term [20], locality constraint SRC [21], regularized robust coding [22], extended SRC [23], superposed SRC [24], sparse- and dense- hybrid representation [25], SRC based feature extraction [26], and dictionary learning for SRC [27]–[29].

If images originate from distinct subjects are less correlated than those from the same subject, a sparse representation is expected with the significant representation coefficients focusing on the correct subject. However, the subspaces spanned by different subjects do have some correlation due to their similar appearances. Moreover, images captured under the same condition but from different subjects may show higher correlation than those from the same subject under different conditions. Therefore, the unsupervised l_1-norm minimization

Manuscript received February 5, 2015; revised July 8, 2015 and February 2, 2016; accepted March 5, 2016. Date of publication March 22, 2016; date of current version May 23, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Weisi Lin.

The authors are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jlai1@ntu.edu.sg; exdjiang@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2545249

of the sample coefficients in SRC may result in the representation of a query image by samples from many different subjects, which may cause misclassification. To alleviate this problem, it is a rational way to harness the label information of training data [27], [30], [31]. Similarly, Group Lasso (GL) [32] is proposed to seek an image representation with minimized $l_{2,1}$-norm of the sample coefficients. However, the $l_{2,1}$-norm not only minimizes the number of the selected subjects, but also minimizes the l_2-norm of coefficients within each class. The later may prevent the attained result from the desired solution as the optimal representation of a query image by training samples of the correct subject may not necessarily be dense.

The first part of this work aims at alleviating the problems of SRC and GL. Towards this end, we propose a new algorithm, named class-wise sparse representation (CSR). Instead of minimizing the number of training samples used in SRC, it minimizes the number of training classes in representing the query image. Different from the GL, the proposed CSR has more emphasis on the sparsity between the classes and hence more likely selects the correct class. In implementation, the desired solution can be obtained by minimizing the representation error and the class-wise sparsity simultaneously, which can be formulated as a convex optimization problem solved via Augmented Lagrange Multiplier scheme [33], [34].

The holistic approaches have been demonstrated being effective for well controlled training and testing data. However, they are vulnerable to the uncontrolled data (e.g., extreme shadows, expression, and disguise) [25], [35], [36]. Local approaches, such as keypoint based [37]–[40], component based [41]–[43] and patch based [15], [16], [44], [45] methods, are shown to be more robust in dealing with the extreme variation. However, both keypoint and component based methods have huge computational burden and their good performances heavily depend on the reliable detection of the keypoints or components, which is not an easy task. Patch based approaches that simply partition the image into predefined patches are often preferred for their usability. Conventional patch based approaches apply the classifier to each patch separately and fuse the patch classification results for final decision. Many fusion methods have been studied, such as product rule, sum rule, max rule, median rule, and weighted sum rule. In [15], LRC is applied to each patch and the label of the query image follows the result of patch that has the minimum representation error. Differently, SRC by patch [16] applies the majority voting to the classification results of individual patches. It treats all patches equally regardless whether they are corrupted or not. To alleviate this problem, modular weighted global sparse representation (WGSR) [45] discards the corrupted patches based on the patch residual and sparsity, and uses the remaining patches for classification by weighting patches with a nonlinear function. However, the optimization of the weighting function remains an open issue.

As each patch is processed separately, patch based approach is more flexible than holistic one. Without the bounding between patches, the corrupted patches caused by extreme variation will not affect the representation and classification of the representative ones. Unfortunately, the patch scheme largely reduces the discriminative information used in the representation and classification. The classification result based on the single patch is much less reliable than that from the whole image. Fusing these less reliable results does not guarantee a correct classification.

To exploit the flexible representation of patches and alleviate the unreliability problem caused by the separated patch optimization, a collaborative patch (CP) representation scheme, together with the proposed CSR, named CSR-CP, is proposed in the second part of this paper. Different from the holistic approaches that restrict the representation coefficient being the same for all pixels of a training image, each patch in the proposed scheme has its own representation coefficient. Different from the aforementioned patch based approaches that optimize each patch representation separately, the CP scheme puts all patches of an image into a group, and optimizes a group-wise sparse representation of the whole query image. By seeking the group-wise sparsity over the coefficients of all patches, it harnesses the relationship among patches of the same image to obtain a reliable and discriminative result. The proposed CP scheme can be realized by extending the proposed CSR algorithm to the group of patches. Comprehensive experimental results are presented in this paper to verify the effectiveness of the proposed CSR-CP approach on several benchmark databases. They show that it significantly outperforms the sparse representation related holistic and patch based approaches.

II. Class-Wise Sparse Representation

An image of m pixels is arranged in a column vector. Let $\mathbf{A}_i = [\mathbf{a}_{i,1}, \ldots, \mathbf{a}_{i,n_i}] \in \Re^{m \times n_i}$ be a stack of n_i training samples with dimensionality of m from the ith subject. $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_C] \in \Re^{m \times n}$ stacks training samples of all C subjects and $n = \sum_{i=1}^{C} n_i$. A query image $\mathbf{y} \in \Re^m$ can be represented by training samples of all classes $\mathbf{A}$ as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{x} \in \Re^n$ is a representation coefficient vector associated with $\mathbf{A}$, and $\mathbf{e}$ is the representation error. As it is in general assumed that samples of a specific subject lie in a linear subspace, a query image is expected to be well represented as a linear combination of the training samples of the same class. Thus, we can find a sparse solution to the representation coefficients $\mathbf{x}$. By measuring the sparsity of the coefficient vector with l_1-norm, the sparse optimization problem is formulated as

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{2}$$

where $\|\cdot\|_1$ denotes the l_1-norm that is absolute sum of a vector and λ is a parameter for compromise between the reconstruction error and sample-wise sparsity. Many algorithms are proposed to solve problem (2) efficiently [46]–[48]. With the assumption that subspaces of distinct classes are independent to each other, (2) achieves a discriminative representation where significant nonzero coefficients are only associated to the correct subject [49], [50]. Therefore, sparse representation (2)

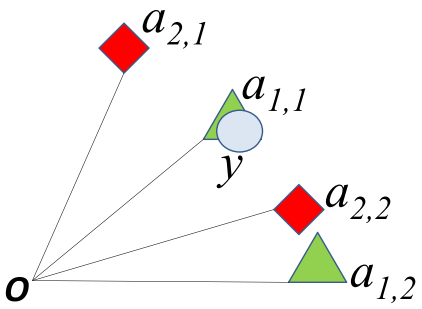


Fig. 1. An example of two different representations: a query image $\mathbf{y}$ (circle) can be sparsely well represented by samples $\mathbf{a}_{1,1}$ and $\mathbf{a}_{1,2}$ of one class (triangle); it can also be densely well represented by samples $\mathbf{a}_{2,1}$ and $\mathbf{a}_{2,2}$ of another class (diamond).

is applied to classification tasks, such as face recognition, and named sparse representation based classification (SRC) scheme [16].

However, there are two problems of SRC for face recognition: First, the significant nonzero coefficients, though sparse, could be associated with many subjects. Due to the common facial components, the assumption of independence of face subspaces cannot be always guaranteed, which leads to significant nonzero coefficients spreading into many distinct subjects. Second, to achieve a sparse coefficient, SRC tends to randomly select a single representative from the highly correlated training samples [51], [52]. The facial images have not only the identity information but also much other information, such as illumination and expression. Thus, the correlation between the samples of some different subjects with similar variations could be higher than those of the same subject with different variations. In this case, SRC is known to have stability problems [52]. Typically, given two highly correlated samples, SRC will randomly select one of the two [52]. The randomness of SRC may lead to an unreliable result.

Although these two problems seem unrelated, both of them can be resolved by using the class labels of training data during sparse optimization. With this discriminative information, the significant representation coefficients can be forced to associate with few subjects. Selecting classes instead of individual samples can alleviate the problem of random selection of highly correlated data. Therefore, Group Lasso (GL) [32] is proposed to conquer the problems of SRC by replacing the l_1-norm with the so-called $l_{2,1}$-norm as following

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{C} \|\mathbf{x}_i\|_2, \tag{3}$$

where $\mathbf{x}_i \in \Re^{n_i}$ is the representation coefficient vector of the ith class and $\mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_C]$. The second term of (3) is $l_{2,1}$-norm [53] that is widely used for group sparsity measurement.

The $l_{2,1}$-norm is in fact the combination of a l_2-norm within class and a l_1-norm across classes. It is not difficult to understand the role of l_1-norm minimization over classes, which enforces a group sparse representation. However, the minimization of l_2-norm suppresses significant sparse coefficients and promotes a dense representation within the class. Take Fig. 1 as an example, where four training samples (i.e., $\mathbf{a}_{1,1}$, $\mathbf{a}_{1,2}$, $\mathbf{a}_{2,1}$, $\mathbf{a}_{2,2}$) of two subjects (triangle and diamond) and a query sample $\mathbf{y}$ (circle) are shown. The query sample can be well represented by the samples of the triangle subject, i.e., $\mathbf{y} = 0.97 \cdot \mathbf{a}_{1,1} + 0.01 \cdot \mathbf{a}_{1,2}$. It can also be well represented by a combination of samples of the diamond subject as $\mathbf{y} = 0.5 \cdot \mathbf{a}_{2,1} + 0.5 \cdot \mathbf{a}_{2,2}$. Although the l_1-norm of the former is smaller than the later, its l_2-norm (0.970) is much larger than the later (0.707). Therefore, the l_2-norm minimization of GL results in the death of the single large coefficient and the survival of the two small weights. The former one, however, is more likely to be the correct subject because the face image variation is more likely sparsely distributed rather than a dense Gaussian distribution. Therefore, the l_2-norm minimization in the second term of (3) adversely affects the correct class selection during the minimization process of GL.

To tackle the problems of SRC and GL, we propose the following optimization problem:

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{z}\|_0 \text{ s.t. } z_i = \|\mathbf{x}_i\|_2,$$
$$\text{where } \mathbf{z} = [z_1, z_2, \ldots, z_C]^T. \tag{4}$$

Here the first term is the reconstruction error, the second term is a class-wise sparsity measurement, and λ is a parameter balancing the effects of these two terms.

The minimization of l_0-norm over $\mathbf{z}$ in (4) is a non-convex problem, which can only be solved by the exhaust searching [54]. Fortunately, recent studies in compress sampling [55]–[57] show that, the sparse vector can be approximately recovered by replacing the non-convex l_0-norm with the convex l_1-norm if the solution is sparse enough. To be robust to the gross images, which is ubiquitously contaminated by sparse noises with arbitrary magnitudes [58] and therefore cannot be well characterized by the l_2-norm, l_1-norm is used to measure the representation error. As a result, we use l_1-norm minimization for both sparse reconstruction error and class-wise sparsity. Therefore, we propose a class-wise sparse representation (CSR) method as

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_1 + \lambda\|\mathbf{z}\|_1 \text{ s.t. } z_i = \|\mathbf{x}_i\|_2,$$
$$\text{where } \mathbf{z} = [z_1, z_2, \ldots, z_C]^T. \tag{5}$$

The proposed CSR introduces a new variable $\mathbf{z}$ into the optimization process. The minimization of the number of nonzero elements of the C dimensional vector $\mathbf{z}$ directly coincides with the aim of finding the most class-wise sparse representation of C classes. The regularization term in (5) searches sparse z_i, $i = 1, 2, \ldots, C$ in $\Re^C$ space. In contrast, the regularization term in (3) searches a group sparse representation x_k, $k = 1, 2, \ldots, n$ in $\Re^n$. Its minimization of the l_2-norm of $\mathbf{x}_i$ may lead to an undesirable result as the optimal representation of a query image by training samples of the correct subject may not necessarily be dense. In addition, it is troublesome to minimize the l_2-norm of the representation error if there are sparse but large corruptions in images. By characterizing the representation error with the l_1-norm in the proposed CSR, the adversarial impact of large noise is suppressed. The

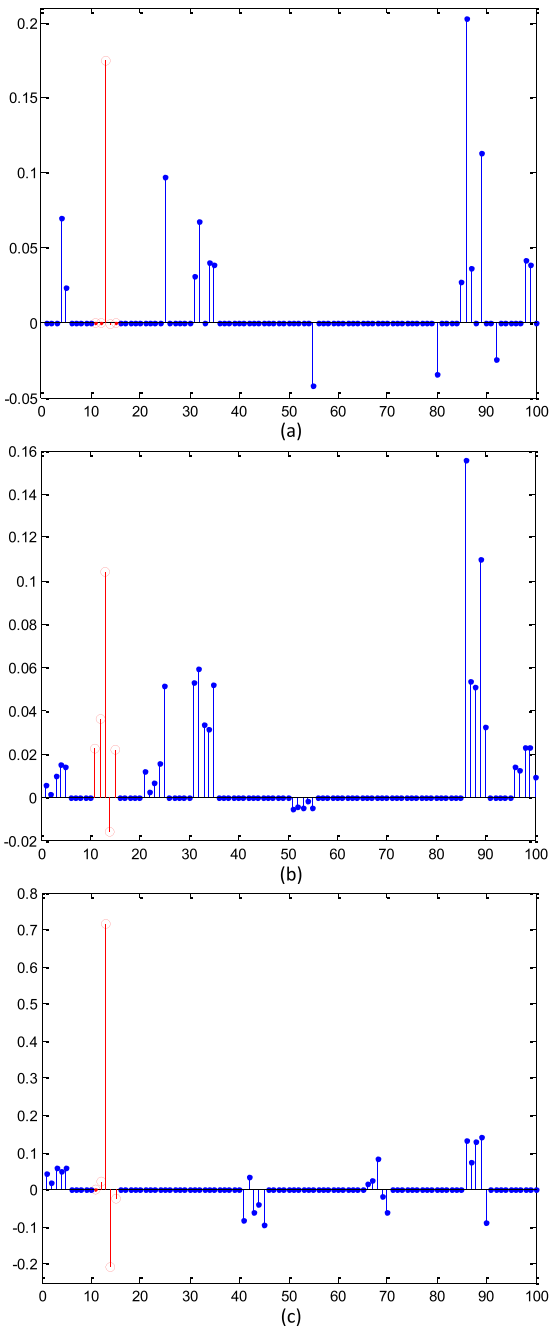


Fig. 2. Comparison of representation coefficients of SRC (a), GL (b), and the proposed CSR (c) on AR database of 20 persons with 5 undisguised training images per person. The x-axis is $5(k-1)+j$ where k and j are respectively the indices of subjects and training images of a subject. Coefficients of the correct subject are in red lines (circles) and others are in blue lines (dots).

two new components in the proposed CSR (5) lead to the development of new optimization algorithm, which will reach different results as shown in the experiments. In some cases, the performance improvement of the proposed CSR (5) over group Lasso (3) is significant.

Fig. 2 visualizes a comparison of coefficients of SRC, GL, and the proposed CSR. Fig. 2a shows that SRC achieves the most sample sparse coefficients with only 17 nonzero elements. However, they spread into 9 different subjects as it does not use the class label information of training data. The most significant coefficient is associated with the wrong subject that leads to misclassification in this example. Fig. 2b shows coefficients of GL, in which 7 subjects have nonzero coefficients. The l_2-norm leads to a dense representation within the class. The correct class fails to win the most significant representation coefficient, which results in the wrong decision. For the proposed CSR, it achieves the most class-wise sparse coefficients (5 classes) among the three methods. Furthermore, it makes the correct classification by assigning the correct subject with the largest coefficient, which stands out against the others. Fig. 2 shows that the proposed CSR achieves the most discriminative representation among the three.

A. Optimization

To solve problem (5), we first convert it to the following equivalent optimization problem by introducing two auxiliary variables $\mathbf{e} \in \Re^m$ and $\mathbf{u} \in \Re^n$:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z},\mathbf{e},\mathbf{u}} \frac{1}{2}\|\mathbf{e}\|_1 + \lambda\|\mathbf{z}\|_1 \\ &\text{s.t. } \mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x},\ \mathbf{x} = \mathbf{u},\ z_i = \|\mathbf{u}_i\|_2, \end{aligned} \tag{6}$$

where $\mathbf{u}_i \in \Re^{n_i}$ is a sub-vector of $\mathbf{u}$ with coefficients associated with class i.

For efficiency, we adopt the Augmented Lagrange Multiplier scheme [33], [34], which has been successfully applied to a variety of convex or nonconvex problems [53]. Consequently, it derives the following unconstrained optimization problem:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z},\mathbf{e},\mathbf{u}} &\frac{1}{2}\|\mathbf{e}\|_1 + \lambda\|\mathbf{z}\|_1 \\ &+ \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + \boldsymbol{\theta}^T(\tilde{\mathbf{u}} - \mathbf{z}) + \boldsymbol{\beta}^T(\mathbf{x} - \mathbf{u}) \\ &+ \frac{\mu}{2}\left(\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 + \|\tilde{\mathbf{u}} - \mathbf{z}\|_2^2 + \|\mathbf{x} - \mathbf{u}\|_2^2\right), \end{aligned} \tag{7}$$

where $\boldsymbol{\gamma} \in \Re^m, \boldsymbol{\theta} \in \Re^C$ and $\boldsymbol{\beta} \in \Re^n$ are the Lagrangian multipliers, $\mu > 0$ is the penalty parameter, and $\tilde{\mathbf{u}} \in \Re^C$ denotes the vector $\left[\|\mathbf{u}_1\|_2, \|\mathbf{u}_2\|_2, \ldots, \|\mathbf{u}_\mathbf{C}\|_2\right]^T$.

Instead of optimizing all variables simultaneously, as they are separable, we apply the alternating optimization scheme, which has been widely advocated by sparse and low-rank optimization works [25], [50], [59], [60]. As a result, the original problem is decomposed into several subproblems, which optimize one variable by fixing the others in each step.

Given the current $\mathbf{x}, \mathbf{z}$ and $\mathbf{u}$, we optimize $\mathbf{e}$ by solving the following subproblem:

$$\begin{aligned} &\arg\min_{\mathbf{e}} \frac{1}{2}\|\mathbf{e}\|_1 + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 \\ &= \arg\min_{\mathbf{e}} \frac{1}{2\mu}\|\mathbf{e}\|_1 + \frac{1}{2}\|\mathbf{e} - (\mathbf{y} - \mathbf{A}\mathbf{x} + \boldsymbol{\gamma}/\mu)\|_2^2. \end{aligned} \tag{8}$$

Problem (8) can be solved via the soft-thresholding operator [61].

We fix the other variables and optimize $\mathbf{x}$ by minimizing the following problem:

$$\begin{aligned}\arg\min_{\mathbf{x}} &\boldsymbol{\gamma}^T(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{e})+\boldsymbol{\beta}^T(\mathbf{x}-\mathbf{u})\\ &+\frac{\mu}{2}(\|\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{e}\|_2^2+\|\mathbf{x}-\mathbf{u}\|_2^2)\\ =\arg\min_{\mathbf{x}} &\frac{\mu}{2}\mathbf{x}^T(\mathbf{A}^T\mathbf{A}+\mathbf{I})\mathbf{x}\\ &-\mu\big((\mathbf{y}-\mathbf{e}+\boldsymbol{\gamma}/\mu)^T\mathbf{A}+(\mathbf{u}-\boldsymbol{\beta}/\mu)^T\big)\mathbf{x}. \end{aligned} \tag{9}$$

This is a least square problem, which leads to a close-form solution of $\mathbf{x}$ as:

$$\mathbf{x}=(\mathbf{A}^T\mathbf{A}+\mathbf{I})^{-1}\big(\mathbf{A}^T(\mathbf{y}-\mathbf{e}+\boldsymbol{\gamma}/\mu)+(\mathbf{u}-\boldsymbol{\beta}/\mu)\big). \tag{10}$$

With the updated $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{e}$, the optimized $\mathbf{u}$ can be obtained by

$$\begin{aligned}\arg\min_{\mathbf{u}} &\boldsymbol{\theta}^T(\tilde{\mathbf{u}}-\mathbf{z})+\boldsymbol{\beta}^T(\mathbf{x}-\mathbf{u})\\ &+\frac{\mu}{2}(\|\tilde{\mathbf{u}}-\mathbf{z}\|_2^2+\|\mathbf{x}-\mathbf{u}\|_2^2). \end{aligned} \tag{11}$$

By simple manipulation, (11) can be rewritten as

$$\arg\min_{\mathbf{u}}\sum_{i=1}^{C}\Big[(\theta_i-\mu z_i)\|\mathbf{u}_i\|_2+\mu\|\mathbf{u}_i-(\mathbf{x}_i+\boldsymbol{\beta}_i/\mu)/2\|_2^2\Big], \tag{12}$$

which has a closed form solution by the 1D shrinkage formula [53]:

$$\mathbf{u}_i=\max\Big(\|\mathbf{r}_i\|_2-\frac{(\theta_i-\mu z_i)}{2\mu},0\Big)\frac{\mathbf{r}_i}{\|\mathbf{r}_i\|_2}, \tag{13}$$

where $\mathbf{r}_i=(\mathbf{x}_i+\boldsymbol{\beta}_i/\mu)/2$, for $i=1,2,\ldots,C$.

We can acquire the optimum $\mathbf{z}$ by solving the following problem:

$$\begin{aligned}\arg\min_{\mathbf{z}} &\lambda\|\mathbf{z}\|_1+\boldsymbol{\theta}(\tilde{\mathbf{u}}-\mathbf{z})+\frac{\mu}{2}\|\tilde{\mathbf{u}}-\mathbf{z}\|_2^2\\ =\arg\min_{\mathbf{z}} &\frac{\lambda}{\mu}\|\mathbf{z}\|_1+\frac{1}{2}\|\mathbf{z}-(\tilde{\mathbf{u}}+\boldsymbol{\theta}/\mu)\|_2^2. \end{aligned} \tag{14}$$

The same as problem (8), (14) can be solved by the soft-thresholding operator.

Finally, the Lagrangian multipliers are updated as

$$\boldsymbol{\gamma}=\boldsymbol{\gamma}+\mu(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{e}), \tag{15}$$

$$\boldsymbol{\theta}=\boldsymbol{\theta}+\mu(\tilde{\mathbf{u}}-\mathbf{z}), \tag{16}$$

$$\boldsymbol{\beta}=\boldsymbol{\beta}+\mu(\mathbf{x}-\mathbf{u}). \tag{17}$$

The steps (8), (9), (12), and (14)-(17) are repeated until the convergence conditions are attained. Algorithm 1 summarizes the procedures of solving the optimization problem (5). In (12), different from GL using a constant weight for all $\|\mathbf{u}_i\|_2$, CSR varies the weights from class to class. The weights play an important role in the optimization. It is easy to see in (14) that larger $\|\mathbf{u}_i\|_2$ leads to larger z_i. As shown in (12), larger z_i decreases the weight of $\|\mathbf{u}_i\|_2$, which will less penalize the l_2-norm of $\mathbf{u}_i$. Therefore, this makes the proposed CSR emphasize the sparsity of $\|\mathbf{u}_i\|_2$ over i and deemphasize the minimization of the l_2-norm of individual $\mathbf{u}_i$. Furthermore, compared to the proposed CSR, GL lacks two crucial steps (8) and (14). In (8), the proposed CSR attempts to find the sparse noise vector $\mathbf{e}$. It is well known that the l_2-norm minimization (9) is not robust to outliers, especially, when the outliers are significant different from the original data. By removing the sparse but arbitrary noise via (8), the optimized $\mathbf{x}$ in (9) will less overfit to a corrupted query sample $\mathbf{y}$. In (14), we minimize the l_1-norm of $\mathbf{z}$ by searching z_i and setting the small nonzero elements of $\tilde{\mathbf{u}}$ to zero. As a result, the proposed CSR achieves more class-wise sparse coefficients compared to GL, which is further verified in the experiments.

Algorithm 1 Class-Wise Sparse Representation (CSR)

Input: Matrixes of training sample set $\mathbf{A}$, query image $\mathbf{y}$, parameter λ.

Initialization: $\mathbf{x}=\mathbf{0},\mathbf{z}=\mathbf{0},\mathbf{u}=\mathbf{0},\boldsymbol{\gamma}=\mathbf{0},\boldsymbol{\theta}=\mathbf{0},\boldsymbol{\beta}=\mathbf{0},\mu=10^{-3},\mu_{max}=10^6,\rho=1.1,\epsilon=10^{-3}$.

while *not converged* **do**

1. update the $\mathbf{e}$ as Problem (8).
2. update the $\mathbf{x}$ as Problem (9).
3. update the $\mathbf{u}$ as Problem (12).
4. update the $\mathbf{z}$ as Problem (14).
5. update the multipliers as (15)-(17).
6. update the parameter μ by $\mu=min(\rho\mu,\mu_{max})$.
7. check the convergence conditions: $\|\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{e}\|_\infty<\epsilon$ and $\|\tilde{\mathbf{u}}-\mathbf{z}\|_\infty<\epsilon$ and $\|\mathbf{x}-\mathbf{u}\|_\infty<\epsilon$.

end

Output: $\mathbf{x}$, $\mathbf{e}$.

B. Classification

Once the coefficient vector $\mathbf{x}$ and representation error $\mathbf{e}$ are obtained, the corrupted query image $\mathbf{y}$ can be recovered as

$$\hat{\mathbf{y}}=\mathbf{y}-\mathbf{e}=\mathbf{A}\mathbf{x}. \tag{18}$$

Now the reconstruction error by a single class, $\mathbf{e}_i$, can be generated as

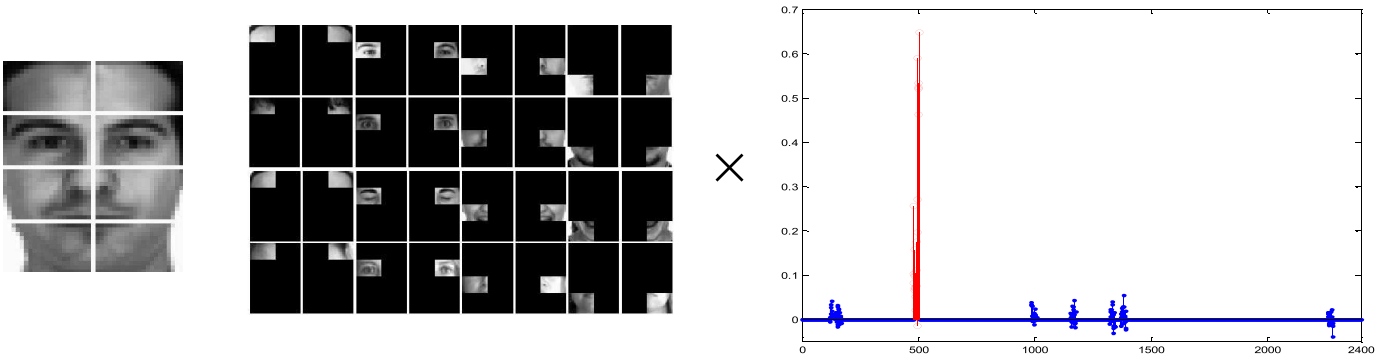
$$\mathbf{e}_i=\hat{\mathbf{y}}-\mathbf{A}\delta_i(\mathbf{x}), \tag{19}$$

where $\delta_i(\mathbf{x})$ only retains the coefficients associated with class i and the others are 0.

The reconstruction by samples of class i will be the closest to $\hat{\mathbf{y}}$ if the query image originates from the same class. Thus, the reconstruction error of the correct class should be the minimum among all. To measure the magnitude of class reconstruction error, l_2-norm is widely applied. However, the quadratic term of l_2-norm makes it easily disturbed by a few or even a single large outlier. To tackle this problem, l_1-norm is a popular choice in field of robust statistics. In this paper, we propose to measure the magnitude of class error as

$$r_i=\|\mathbf{e}_i\|_1=\|\mathbf{A}\mathbf{x}-\mathbf{A}\delta_i(\mathbf{x})\|_1=\|\mathbf{A}\bar{\delta}_i(\mathbf{x})\|_1, \tag{20}$$

where $\bar{\delta}_i(\mathbf{x})\in\Re^n$ is a vector generated by setting the elements of $\mathbf{x}$ associated with the ith class to 0.



(a) (b)

Fig. 3. (a) A query image partitioned into the predefined patches. (b) An example of collaborative patch sparse representation of the whole query image. The coefficients indicated by red lines (circle) are associated with the correct class.

Finally, the query image is labeled to the subject with the minimum residual

$$i^* = \arg\min_i r_i. \tag{21}$$

III. Collaborative Patch Representation

SRC [16] has demonstrated its effectiveness in face recognition under some scenarios. However, a violation of carefully controlled training samples may result in severe performances degradation. For example, to represent a normal query image, instead of choosing the extremely illuminated training samples of the correct subject, SRC tends to select the normal samples from other subjects. In this case, the result of SRC is no longer informative.

In order to handle face with extremely variation (e.g., shadow, expression, and disguise), in [16], images are partitioned into L patches as shown in Fig. 3a. It generates a set of patch dictionaries $\mathbf{A}^l \in \Re^{(m/L)\times n}$ by stacking the lth patches of n training images, for $l = 1, 2, \ldots, L$. Similarly, a query image is also partitioned into $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^L \in \Re^{m/L}$. For each patch, $\mathbf{y}^l$ is represented by the patch dictionary $\mathbf{A}^l$ as

$$\mathbf{y}^l = \mathbf{A}^l\mathbf{x}^l + \mathbf{e}^l, \tag{22}$$

where $\mathbf{x}^l \in \Re^n$ and $\mathbf{e}^l \in \Re^{m/L}$ are the patch representation coefficients and error, respectively. For each of L patches, SRC (2) is applied to (22) for classification. The label of query image is decided based on the majority voting of all patches [16]. In this paper, we call this method SRC by patch (SRC-P).

SRC-P alleviates some limitations of SRC. By partitioning the image, the corrupted patches caused by extreme variation will not affect the representation and classification of the representative ones. Moreover, for the representation of each patch, SRC-P solves a more underdetermined linear system of $(m/L)\times n$, which has only one Lth constraints of SRC $(m\times n)$, and hence it reconstructs the query image more accurately.

In the procedure of SRC-P, it searches individual local optimums first (i.e., the result of each patch) and then fuses them for a final decision. A patch with the dimensionality of m/L has much less discriminative information than the whole image with the dimensionality of m. Based on the reduced discriminative information, the local optimums of SRC-P are less reliable. The fusion of these less reliable results does not guarantee a correct classification. Furthermore, as all patches are processed separately, they are not constrained to select the same classes. Thus, the significant representation coefficients of different patches could be distributed in a number of different training classes with no class wining more patches than the others.

To alleviate these problems, we propose a scheme that gathers all patches into a single joint optimization. Given a training image $\mathbf{a}_{i,k}$, we generate L patch images, $\mathbf{a}^1_{i,k}, \mathbf{a}^2_{i,k}, \ldots, \mathbf{a}^L_{i,k} \in \Re^m$, where the only nonzero elements of $\mathbf{a}^l_{i,k}$ are the pixels of $\mathbf{a}_{i,k}$ associated with the lth patch. Some examples of $\mathbf{a}^l_{i,k}$ are shown in the left image of Fig 3b, in which the patch images of a row are generated by an image. We stacks all L patch images of $\mathbf{a}_{i,k}$ as $\mathbf{A}^P_{i,k} = [\mathbf{a}^1_{i,k}, \mathbf{a}^2_{i,k}, \ldots, \mathbf{a}^L_{i,k}]$. $\mathbf{A}^P_i = [\mathbf{A}^P_{i,1}, \mathbf{A}^P_{i,2}, \ldots, \mathbf{A}^P_{i,n_i}]$ encloses all patch images of a subject. By concatenating the patch images of all C subjects, a dictionary is produced as $\mathbf{A}^P = [\mathbf{A}^P_1, \mathbf{A}^P_2, \ldots, \mathbf{A}^P_C] \in \Re^{m\times nL}$. We propose a collaborative patch (CP) representation of query sample $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{A}^P\mathbf{x}^P + \mathbf{e}^P, \tag{23}$$

where $\mathbf{x}^P \in \Re^{nL}$ and $\mathbf{e}^P \in \Re^m$ are the collaborative patch representation coefficients and error, respectively.

It is not difficult to see that if the proposed CSR (5) is applied to solve $\mathbf{x}^P$ in (23), the representation coefficients of all patches of a training image are put into a group to compete with others in the sparse optimization. Thus, all patches of an image collaboratively participate in the optimization. Therefore, we propose the CP representation scheme together with the proposed CSR, named CSR-CP, as following optimization problem

$$\min_{\mathbf{x}^P, \mathbf{z}^P} \frac{1}{2}\|\mathbf{y} - \mathbf{A}^P\mathbf{x}^P\|_1 + \lambda\|\mathbf{z}^P\|_1 \text{ s.t. } z^P_i = \|\mathbf{x}^P_i\|_2$$
$$\text{where } \mathbf{z}^P = [z^P_1, z^P_2, \ldots, z^P_C]^T, \tag{24}$$

here $\mathbf{x}^P_i \in \Re^{n_i L}$ is the coefficient vector associated with $\mathbf{A}^P_i$. All patches in $\mathbf{A}^P$ and $\mathbf{y}$ are normalized to unit l_2-norm. Different from SRC-P, which considers each patch representation separately, the proposed CSR-CP method bounds all

patches of an image and all images of a class into a group. Putting all patches of an image together in the optimization, CSR-CP compensates the loss of the discriminative information caused by partition of the image into patches. Thus, the result of the proposed CSR-CP is more informative than SRC-P for classification.

Once (24) is solved by Algorithm 1, a classification procedure similar to (20) and (21) is applied.

SRC constrains all pixels of an image to share the same coefficient. Thus, it either selects all or none pixels of an image. SRC-P attempts to interpolate between these two conditions by partitioning image into patches and processing them separately. It results in that every patch has its own representation coefficient. That is, the representative patches of the corrupted image can be freely and independently used to reconstruct those of query image. Therefore, SRC-P shows its effectiveness when the training data is not well controlled, such as with disguise, extreme expression and shadow. Although SRC-P can achieve a more accurate representation, it does not necessarily result in better recognition rates in all cases. In fact, SRC-P leads the representation to an extreme, which does not take the relations between patches of the same image into consideration. Hence, SRC-P largely reduces the discriminative information, which is critical for classification. The proposed CSR-CP makes a compromise between SRC that treats the whole image as an entirety, and SRC-P that allows different representations for different patches. It uses all patches collaboratively to learn a group-wise sparse representation where patches are united as an entirety in the sparse optimization. The proposed CSR-CP, which inherits some merits of both holistic and patch based approaches, attains a more reliable and robust representation.

To illustrate a problem the proposed approach solves, Fig. 4 shows a real example comparing CSR-CP with SRC and SRC-P. An undisguised query image is represented by training samples that contains 100 subjects, each of which has two images, from AR databases. The two training images of the correct subject are both disguised, one wearing sunglasses and the other wearing scarf. SRC-P and the proposed CSR-CP apply 8 patches as shown in Fig. 3. The representation accuracy $(1 - r_i / \max(r_i))$ by training samples of a single class of SRC, SRC-P and the proposed CSR-CP are shown in Fig. 4a, b and c, respectively. Due to the disguise in training samples of the correct subject, SRC tends to select the uncorrupted samples of many other subjects to represent the query image. As a result, there are 4 class-wise representation accuracies higher than that of the correct subject, which leads to wrong classification. As SRC-P optimizes patches separately, the different patches have different selections of subjects. This can be seen in Fig. 4b that many subjects have nonzero class-wise representation accuracies, where 2 subjects have higher class-wise representation accuracies than the correct subject. More specifically, the minimum reconstruction errors of 8 patches are won by 8 different subjects that leads to misclassification. For the proposed CSR-CP method, the correct subject achieves a much higher class-wise representation accuracy than the other subjects and hence it correctly labels the query image.

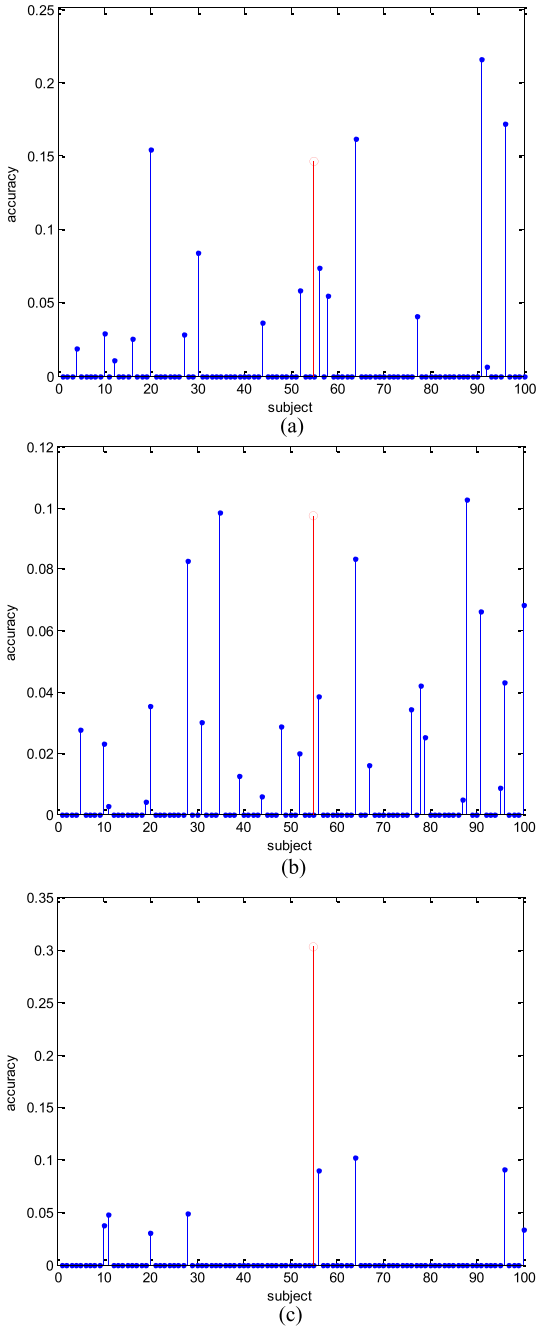


Fig. 4. Comparison of SRC (a), SRC-P (b) and the proposed CSR-CP (c) in terms of representation accuracy by single class. For SRC-P, the class residual r_i is the sum of the 8 patch residuals for comparison with SRC and CSR-CP. The red lines (circle) are associated with the correct subject.

IV. Experiments

The proposed approaches, class-wise sparse representation (CSR) and class-wise sparse representation with collaborative patch (CSR-CP), are evaluated in 4 face databases: Extended Yale B [14], CMU Multi-PIE [62], AR [63], and CMU PIE [64]. To show the effectiveness of the proposed CP separately, we also propose to integrate CP scheme in the Group Lasso [32] named GL-CP. All these three proposed algorithms are compared with LRC [15], SRC [16], Group Lasso (GL) [32], [65], LRC by patch (LRC-P) [15], SRC by patch (SRC-P) [16] and WGSR [45]. The parame-

Fig. 5. Face samples from the Extended Yale B.

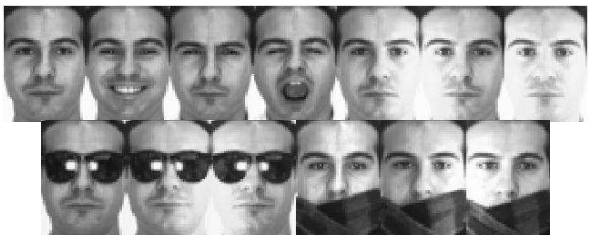


Fig. 6. Face samples from the AR database.

ters λ of the proposed methods are determined by cross-validation on AR dataset and fixed over all experiments of this paper to 0.25 and 0.1 for CSR (5) and CSR-CP (24), respectively, though better performances may be achieved if they are fine tuned to fit each specific experiment. For the competing methods, their parameters are tuned to achieve the best performances on AR dataset. The grey level of image is used as the feature. In all but Experiment 7, the number of patch is fixed to 8 as shown in Fig. 3, which is the same as [15], [16], and [45].

The cropped Extended Yale B database includes 38 subjects, each of which has 64 frontal face images captured under 64 different illuminations. These 64 samples are divided into 5 subsets depending on the angle between the directions of light and face. Fig. 5 shows some samples with various lighting condition. The image size is downsampled from 198×168 to 48×42.

The CMU Multi-PIE database contains face images captured in 4 sessions with variations in illumination, expression and pose. We use the frontal images with neural expression of the first 105 subjects of Session 1 in the experiments. Images are cropped based on the eye locations provided by [66] and downsampled to 50×40 pixels.

For AR database, we use a subset including 50 male subjects and 50 female subjects. For each subject, 26 face images are taken in two separate sessions, that is, each session has 13 samples. These two sessions are with the same expression, illumination and disguise variation. However, Session 2 is captured one week later than Session 1. Fig. 6 shows images of a typical session from AR database. All images are resized from 165×120 to 55×40.

The CMU PIE database has 41,368 facial images of 68 subjects. The face images were acquired by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. In the experiment, we choose the images captured from the frontal poses (C27) and use all the images under different illuminations and expressions. As a result, for each of 68 subjects, there are 49 facial images. Some facial images with illumination and expression changes are shown in Fig. 7. All images are with the size of 45×45.

Fig. 7. Face samples from the CMU PIE database.

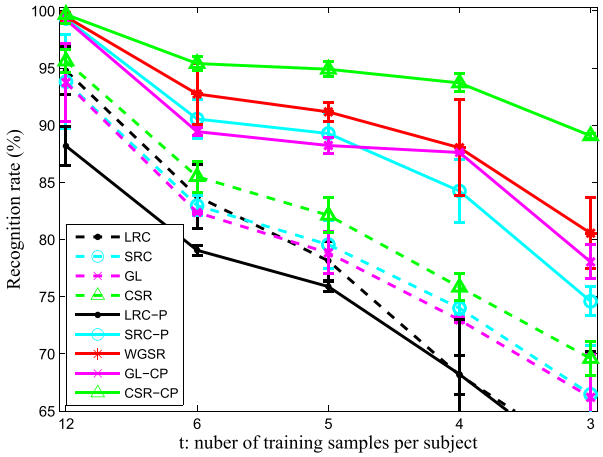


Fig. 8. Recognition Rate on Extended Yale B.

Fig. 9. Face samples with illumination variation from CMU Multi-PIE database.

A. Face Recognition With Uncorrupted Training Samples

This subsection tests the effectiveness of the proposed approaches with uncorrupted training data.

Experiment 1: for Extended Yale B database, we randomly select t from 64 illuminations as training data and the others are used for testing. We repeat this procedure 10 times for each t (12, 6, 5, 4, 3). Fig. 8 plots the averaged results and standard deviations of various methods with different training samples t. Much to our surprise, GL, which attempts to solve the problem of SRC, underperforms SRC in all scenarios. Among the holistic approaches, the proposed CSR achieves the best recognition rates, which is however worse than the patch based methods due to the extreme illuminations and insufficient training data. Although GL underperforms SRC all the time, the proposed GL-CP performs better than SRC-P in the cases of 4 and 3 training samples thanks to the collaborative patch scheme. When t is equal to 12, all patch based methods except for LRC-P achieve close results. WGSR consistently achieves the second best accuracies with the maximum gain over SRC-P 6%. The proposed CSR-CP performs the best and its largest improvement over SRC-P is 14.5%.

Experiment 2: to further verify the performances of various methods in tackling the extreme illuminations, another experiment is done on Multi-PIE database. Following the procedure

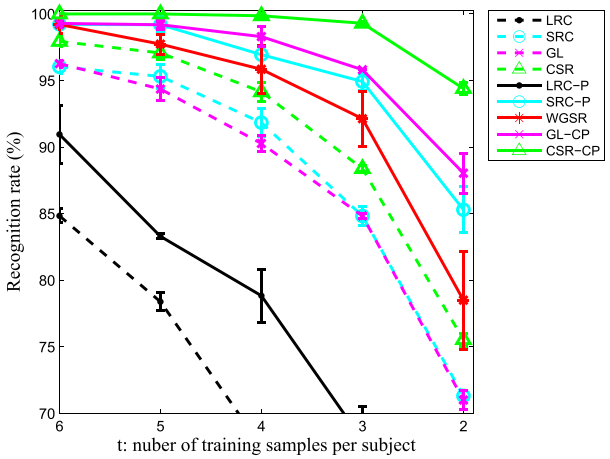


Fig. 10. Recognition rate on Multi PIE Database.

in [62], 18 flash-only images are generated as the differences between flash images (illuminations {1-18}) and non-flash image (illumination {0}) of Multi-PIE database as shown in Fig. 9. For each of the 105 subjects, t images are randomly selected for training and the other $18-t$ images are used as testing data. The averaged recognition rates and standard deviations over 10 runs are plotted against the reduced t in Fig. 10. It can be seen that the proposed CSR visibly outperforms the other holistic approaches. Although the recognition accuracies of GL are similar or even inferior to those of SRC, by using of collaborative patch scheme, the proposed GL-CP outperforms SRC-P and consistently perform the second best. When there are equal or more than 3 training samples per subject, the proposed CSR-CP achieves almost perfect results (i.e., greater than 99%). The maximum accuracy gain of the proposed CSR-CP over SRC-P reaches about 9% at 2 training samples per subject.

Experiment 3: for each subject of AR database, t (6, 5, 4, 3) of 7 undisguised samples from Session 1 are used for training and all 7 undisguised samples from Session 2 are used for testing. With each t, we randomly select 10 times. The averaged results and standard deviations are recorded in Table I. Among all holistic approaches, the proposed CSR achieves the best recognition rates where the performance gain increases with decreasing number of training data. For LRC and SRC, their patch based versions cannot consistently outperform the corresponding holistic ones. In contrast, the proposed CSR-CP and GL-CP perform better than the proposed CSR and GL for all different t, respectively. Again, the proposed CSR-CP consistently achieves the best results.

Experiment 4: for each subject of CMU PIE dataset, t (8, 4, 3, 2) facial images are randomly selected for training and all the rest are used for testing. The averaged recognition rates and standard deviations over 10 runs are shown in Table II. As the variation of this dataset is similar to that of Experiment 3, which includes only small illumination and facial expression, comparable results are obtained. Given the sufficient training data (i.e., $t = 8$), all methods but LRC-P achieve high recognition rates and their differences are marginal. Although SRC-P performs inferior to SRC due

TABLE I
RECOGNITION RATE ON AR FACE DATABASE

number t	6	5	4	3
LRC [15]	72.8±0.47	67.4±1.11	60.5±0.80	53.7±0.85
SRC [16]	91.7±0.07	89.8±0.35	87.5±0.51	81.6±0.88
GL [65]	92.4±0.31	90.3±0.33	88.7±1.09	83.4±0.35
CSR	94.3±0.11	93.7±0.35	93.1±0.33	89.3±0.47
LRC-P [15]	69.2±1.75	67.2±2.05	63.8±2.45	56.5±1.17
SRC-P [16]	92.7±0.70	89.1±0.07	87.7±1.06	79.1±0.25
WGSR [45]	92.7±0.31	90.3±0.64	88.1±0.53	82.4±1.53
GL-CP	94.6±0.44	93.4±0.54	90.7±0.07	86.0±0.71
CSR-CP	**96.7**±0.31	**95.6**±0.44	**94.3**±0.53	**90.8**±0.18

TABLE II
RECOGNITION RATE ON CMU PIE DATABASE

number t	8	4	3	2
LRC [15]	94.0±0.21	76.8±2.40	66.6±2.23	49.6±1.65
SRC [16]	96.2±0.44	92.1±0.42	87.9±0.72	82.4±0.90
GL [65]	95.8±0.29	92.1±0.53	87.8±0.75	82.7±1.04
CSR	96.4±0.15	93.3±0.52	89.8±0.71	85.3±0.73
LRC-P [15]	89.7±0.20	75.2±1.59	69.3±0.37	64.9±4.04
SRC-P [16]	94.7±0.05	90.5±0.42	86.5±0.21	82.4±0.90
WGSR [45]	95.7±0.11	92.3±0.40	87.3±0.30	81.0±1.20
GL-CP	96.0±0.18	92.4±0.23	88.2±0.02	84.5±0.96
CSR-CP	**96.7**±0.14	**94.2**±0.18	**90.3**±0.24	**88.0**±0.70

to the limited discriminative information of each patch, the proposed CSR-CP and GL-CP consistently outperform their corresponding holistic versions. For all number of training data, the proposed CSR-CP and CSR perform as the top 2 approaches.

B. Face Recognition With Occluded Training Samples

This subsection tests the effectiveness of various approaches on training data with disguise.

Experiment 5: for each subject of AR database, all 6 disguised face images (i.e., the second row of Fig. 6) from Session 1 are used for training and 8 face images with illumination variations (i.e., the 1st, 5th, 6th, and 7th samples of the first row of Fig. 6) from both Session 1 and Session 2 are used for testing. Table. III details the performances of various methods in this scenario. The proposed CSR significantly outperforms SRC with the accuracy gain of 9% while that of GL is only 5.4%. The proposed CSR-CP achieves the accuracy gain of 8.1% from SRC-P.

The running time of each approach is evaluated under the Matlab programming environment and on a desktop of 3.5GHZ CPU with 16G RAM. It is 0.02s for LRC and LRC-P, 0.12s for GL, 0.34s for SRC, 0.32s for SRC-P, 0.46s for WGSR, 0.88s for GL-CP, 6.65s for CSR and 19.47s for CSR-CP.

Experiment 6: for each subject of AR database, t (6, 5, 4, 3) facial images are randomly selected from all 13 images of Session 1 for training, and all 13 samples from Session 2 are used for testing. With each t, we repeat this procedure 10 times and record the averaged results and standard deviations in Table IV. It should be noted that the testing data of every subject have two different disguise types, both of which could be absent in its training data but

TABLE III
RECOGNITION RATE ON AR DATABASE WITH ONLY DISGUISED TRAINING SAMPLES

LRC [15]	SRC [16]	GL [65]	CSR	LRC-P [15]	SRC-P [16]	WGSR [45]	GL-CP	CSR-CP
59.2%	83.1%	88.5%	92.1%	75.7%	88.7%	86.2%	93.7%	**96.8**%

TABLE IV
RECOGNITION RATE ON DISGUISED AR FACE DATABASE

number t	6	5	4	3
LRC [15]	48.8±0.46	44.8±1.51	36.7±0.73	27.8±0.50
SRC [16]	78.4±1.54	74.5±0.23	69.0±0.59	59.9±1.43
GL [65]	81.7±1.22	79.0±0.15	72.7±0.69	65.3±0.96
CSR	87.4±1.31	83.7±0.77	80.0±0.98	72.6±1.22
LRC-P [15]	58.9±0.70	58.2±0.13	47.8±3.43	46.5±0.89
SRC-P [16]	79.7±0.31	74.8±1.50	69.9±0.90	60.6±1.77
WGSR [45]	81.7±0.53	78.6±1.13	73.7±0.88	64.9±1.50
GL-CP	82.5±0.93	79.3±1.01	75.4±0.84	69.8±0.64
CSR-CP	**90.4**±1.10	**88.3**±0.99	**84.2**±0.74	**78.9**±1.17

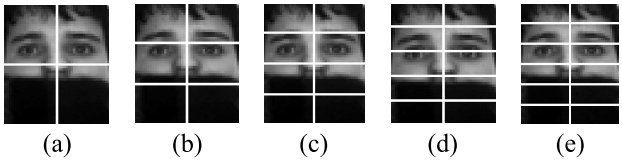


Fig. 11. Face samples partitioned into different numbers of patch. (a) 2×2; (b) 3×2; (c) 4×2; (d) 5×2 (e) 6×2.

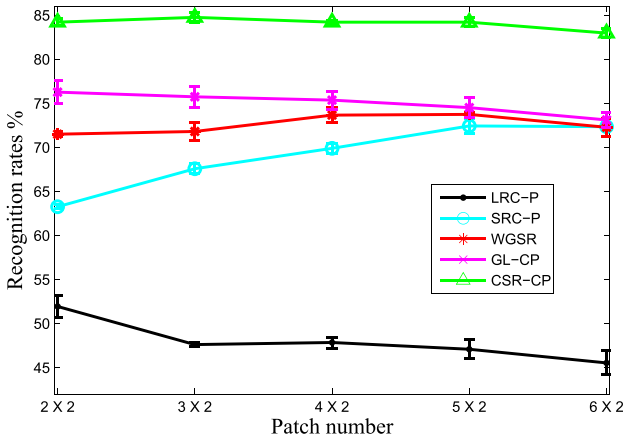


Fig. 12. Recognition rate on AR database with different numbers of patch.

could present in some other subjects. Table IV shows that GL visibly outperforms SRC. Much more significant gains over SRC are achieved by the proposed CSR. The performance of SRC-P is slightly better than that of SRC in this challenge scenario. However, by employing the proposed collaborative patch scheme, the accuracies of both proposed GL-CP and CSR-CP are visibly better than their corresponding holistic ones in all number of training data. The proposed CSR-CP again outperforms the others consistently. The accuracy gain over SRC-P reaches 18.3% at $t = 3$.

C. Face Recognition With Different Patch Numbers

This subsection tests the effectiveness of various approaches with different number of patch L. The images are partitioned into L patches, here L is set to 4 (2×2), 6 (3×2), 8 (4×2), 10 (5×2), or 12 (6×2) as shown in Fig. 11.

Experiment 7: the data setting of Experiments 6 with 4 training samples is used to test various patch based approaches with different number of patch. Fig. 12 plots the performances of averaged results and standard deviations of 10 runs. It shows that the results of SRC-P and LRC-P vary visibly as the number of patch increases. SRC-P delivers recognition rates ranging from 63.3% to 72.4%, while the performance of LRC-P is from 45.5% to 51.9%. The approaches employing relations between patches are less sensitive to the number of patch. The difference of maximum and minimum accuracies of WGSR, the proposed GL-CP and CSR-CP are 2.24%, 3.16% and 1.78%, respectively. Fig. 12 also demonstrates that the proposed CSR-CP significantly outperforms all other algorithms consistently for all numbers of patch.

V. CONCLUSION

SRC considers the image classification problem as a sparse linear representation of the query image. Ideally, all significant representation coefficients should be associated to the correct class. However, as images of different classes could be similar and correlated, the sample-wise sparse representation that ignores the identity information of training samples tends to employ training samples of many subjects. This results in misclassification. GL attempts to solve this problem by minimizing the so-called $l_{2,1}$-norm of the representation coefficients. Although the minimization of the l_1-norm component leads to a group sparsity, the minimization of the l_2-norm component may result in the attained representation deviating from the desired solution as the optimal representation by training samples of the correct subject may not necessarily be dense. The proposed CSR alleviates the problems of SRC and GL. It seeks an optimum representation of the query image by minimizing the number of selected classes of training data.

Patch based approaches are effective to handle the corrupted images and extreme image variations at a price of reducing the discriminative information in the training data. Different from the conventional patch based methods that optimize each patch separately, the proposed CSR-CP approach optimizes all patches together to seek a collaborative patch group-wise sparse representation. It inherits the merits of the patch based approaches in dealing with the corrupted or extremely variate images and alleviates their problems in reducing the discriminative information. As a result, a more reliable and discriminative representation is achieved by the proposed CSR-CP method. Extensive experiments on several benchmark

databases verify the merits and significance of the proposed CSR-CP approach.

REFERENCES

[1] A. Satpathy, X. Jiang, and H.-L. Eng, "Human detection by quadratic classification on subspace of extended histogram of gradients," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 287–297, Jan. 2014.

[2] J. Ren, X. Jiang, and J. Yuan, "Noise-resistant local binary pattern with an embedded error-correction mechanism," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4049–4060, Oct. 2013.

[3] A. Satpathy, X. Jiang, and H.-L. Eng, "LBP-based edge-texture features for object recognition," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1953–1964, May 2014.

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

[5] 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.

[6] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[7] 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.

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

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

[10] X. Jiang and A. H. K. S. Wah, "Constructing and training feed-forward neural networks for pattern classification," *Pattern Recognit.*, vol. 36, no. 4, pp. 853–867, 2003.

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

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

[13] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2003, pp. I-11–I-18.

[14] 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.

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

[16] J. Wright, A. Y. Yang, A. Ganesh, S. 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.

[17] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[18] Y. Liu, F. Wu, Z. Zhang, Y. Zhuang, and S. Yan, "Sparse representation using nonnegative curds and whey," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3578–3585.

[19] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 448–461.

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

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.*, Jun. 2010, pp. 3501–3508.

[28] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1697–1704.

[29] 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.

[30] J. Lai and X. Jiang, "Discriminative sparsity preserving embedding for face recognition," in *Proc. 20th IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3695–3699.

[31] J. Lai and X. Jiang, "Supervised trace lasso for robust face recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

[32] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[33] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York, NY, USA: Academic, 1996.

[34] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," UIUC, Tech. Rep. UILU-ENG-09-2215, 2009

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

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

[37] C. Geng and X. Jiang, "Face recognition based on the multi-scale local image structures," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2565–2575, 2011.

[38] C. Geng and X. Jiang, "Fully automatic face recognition framework based on local and global features," *Mach. Vis. Appl.*, vol. 24, no. 3, pp. 537–549, 2013.

[39] Z. Miao and X. Jiang, "Interest point detection using rank order LoG filter," *Pattern Recognit.*, vol. 46, no. 11, pp. 2890–2901, 2013.

[40] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.

[41] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, vol. 2. Jul. 2001, pp. 688–694.

[42] B. Heisele, P. Ho, J. Wu, and T. Poggio, "Face recognition: Component-based versus global approaches," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 6–21, 2003.

[43] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *Int. J. Comput. Vis.*, vol. 74, no. 2, pp. 167–181, 2007.

[44] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1994, pp. 84–91.

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

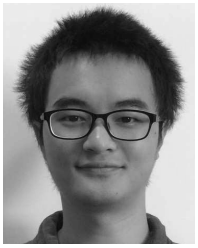
[46] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale ℓ_1-regularized logistic regression," *J. Mach. Learn. Res.*, vol. 8, pp. 1519–1555, Dec. 2007.

[47] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[48] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, "Fast ℓ_1-minimization algorithms for robust face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3234–3246, Aug. 2013.

[49] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2790–2797.

[50] 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. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[51] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B (Statist. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.
[52] E. Grave, G. R. Obozinski, and F. R. Bach, "Trace Lasso: A trace norm regularization for correlated designs," in *Proc. NIPS*, 2012, pp. 1–9.
[53] W. Deng, W. Yin, and Y. Zhang, "Group sparse optimization by alternating direction method," Dept. Comput. Appl. Math., Rice Univ., Houston, TX, USA, Tech. Rep. TR11-06, 2011.
[54] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Comput. Sci.*, vol. 209, nos. 1–2, pp. 237–260, 1998.
[55] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.
[56] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.
[57] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Dec. 2006.
[58] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, May 2011.
[59] 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. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2618–2625.
[60] R. He, W.-S. Zheng, T. Tan, and Z. Sun, "Half-quadratic-based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 261–275, Feb. 2014.
[61] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.
[62] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
[63] A. Martinez and R. Benavente, "The AR face database," Centre de Visio per Computador, Univ. Autonoma Barcelona, Barcelona, CVC Tech. Rep. no. 24, Jun. 1998. [Online]. Available: http://www.cat.uab.cat/Public/Publications/1998/MaB1998/CVCReport24.pdf
[64] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[65] A. Majumdar and R. K. Ward, "Classification via group sparsity promoting regularization," in *Proc. IEEE Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 861–864.
[66] 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.



Jian Lai received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2009, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2015. He is currently a Research Fellow with the School of Computing, National University of Singapore, Singapore. His research interests include pattern recognition, image processing, and computer vision.

Xudong Jiang (M'02–SM'06) received the B.Eng. and M.Eng. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1983 and 1986, respectively, and the Ph.D. degree from Helmut Schmidt University, Hamburg, Germany, in 1997, all in electrical engineering. From 1986 to 1993, he was a Lecturer with 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 Helmut Schmidt University. 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 joined Nanyang Technological University (NTU), Singapore, as a Faculty Member, in 2004, where he served as the Director of the Centre for Information Security from 2005 to 2011. He is currently a Tenured Associate Professor with the School of EEE, NTU. He holds seven patents and has authored over 100 papers with 25 papers in the IEEE journals, including eight papers in the IEEE TRANSACTIONS ON IMAGE PROCESSING, five papers in the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and three papers in the IEEE TRANSACTIONS ON SIGNAL PROCESSING. His research interests include signal/image processing, pattern recognition, computer vision, machine learning, and biometrics. He is an Elected Voting Member of the IFS Technical Committee of the IEEE Signal Processing Society, and serves as an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and *IET Biometrics*.