Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 16, No. 2, pp. 175–194, 2020

IMAGE SET FACE RECOGNITION VIA THE MIXED ℓ_0-NORM SPARSE REPRESENTATION*

Jingjing Liu, Shiwei Ma† and Xianchao Xiu

Abstract: Sparse prior based dictionary learning techniques have been widely applied in the field of computer vision and pattern recognition. Due to the fact that ℓ_0-norm can overcome the drawbacks associated with the ℓ_1-norm, in this paper, an ℓ_0-norm regularized framework with synthesis K-SVD dictionary learning method is proposed. In the first phase, K-SVD method embedded with a fast ℓ_0 algorithm is established to discover an adaptive dictionary for all the training and testing samples. In the second phase, the mixed ℓ_0-norm is embedded into sparse representation classification to regulate the sparsity of inter-class and inner-class level, which has been demonstrated to outperform both SRC(sparse representation classification) and JSRC (joint sparse representation classification). Furthermore, alternating direction method of multipliers is presented to solve our proposed model, and the convergence result is also derived. Finally, we provide extensive comparisons which demonstrate that our method outperforms other state-of-the-arts algorithms on CMU-PIE, YaleB and Multi-PIE databases for multi-view face recognition.

Key words: *sparse prior, K-SVD, mixed ℓ_0-norm, sparse representation classification, joint sparse representation classification, alternating direction method of multipliers*

Mathematics Subject Classification: *90C90, 65K05*

1 Introduction

As more and more digital images are available from a variety of sources, the set to set face recognition (SSFR) has attracted much attention recently [31, 34, 48]. An effective SSFR system should be able to identify face images by computing the similarity of the extracted features in an image set, which may be captured with different poses, illuminations and expressions. In order to achieve good performance in an SSFR system, researchers have proposed different approaches, including the subspace approaches [6,21,45,47], the manifold techniques [13, 36, 37], affine or convex hull [4, 15, 16], the idea based on the nearest points distance [15–17] and dictionary learning (DL) approaches [5, 7, 8, 10, 12, 18, 20, 25, 26, 29, 33, 41, 48].

The core idea behind the DL methods is that the structure information of original images can be efficiently extracted from data itself rather than by mathematical models. Actually, a learnt target dictionary often transcends predefined bases in pattern classification tasks,

*This work was partially supported by the National Natural Science Foundation of China with Grant Nos. (61525401, 61234002, 61872230, 61671285, 61363066, 11671029,51705304) and Natural Science Foundation of Shanghai (Grant No. 19ZR1420800).

†Corresponding author.

benefitting for better flexibility and adaptability to specific data. Therefore, methods for adaptively learning a dictionary with specific properties from instance of data is still an intensively ongoing research in DL. In the field of classification, some methods such as D-KSVD (discrimination K-SVD) [46], LC-KSVD (label consistent K-SVD) [19] have reached quite high accuracies. Previous works mainly emphasize presentative ability more while paying little attention on the potential of discrimination power. Therefore, our main purpose of this paper is to take an exploration for the capability of the Synthesis K-SVD framework in handling pattern classification problems.

It is well known that SRC with embedding ℓ_1-norm algorithm considers classification between each class. Then the development of joint sparse representation based classification (JSRC) was proposed [27, 35], which assumes that the query face images share the same sparsity pattern information with training and general images. Instead of solving the SRC problem for each query image, JSRC with $\ell_{2,1}$-norm is applied to solve a set of query images from the same subject. However, this assumption will not be true when there are large pose differences in the query images. Therefore, forcing the entire view share the same sets of atoms is not applicable for the multi-view face recognition. To overcome this problem, the joint dynamic sparse representation based classification (JDSRC) was proposed [45], and its sparsity brings flexibility to atom selection of JSRC. When the pose variation is large in the query images, JDSRC does not necessarily select the same atom for all poses, this may not lead to a robust solution. In addition, the JDSRC is achieved by an extension of simultaneous orthogonal matching pursuit [35] which is a naive greedy method, Hence, this method may not be convergent. Then, the MSRC [47] is proposed to decrease the influence when a face image has large pose variation in the recognition process. Although the method is convergent, the effect of recognition is not particularly satisfactory. Therefore, a new algorithm is required to solve this challenging multi-view problem.

In [32], the authors argue that the robustness of SRC based methods should be achieved by employing the ℓ_1-norm loss function instead of the ℓ_2-norm loss. As solving this issue is similar to the ℓ_1-norm of Lasso in statistics functional form, i.e., the robust Lasso, which explicitly models the corruptions, is proposed and analysed in [28]. Statistically, this is more generic and proven to be better than the least entropy and error correction alternative discussed in [38]. However, this is achieved at the cost of an extra regularisation parameter. In the related robust paper [30], a slightly different loss function, known as Huber's robust loss function is employed, but it requires the estimates of the Huber's parameters, which induces additional computational burden and over-penalization. The most advanced result is archived by [47], which is introduced a trade off between the ℓ_1-norm from SRC and $\ell_{2,1}$-norm from JSRC to achieve high recognition performance, as when some face images are with a certain degree, this mixed norm will find an optimal representation based on the shared information induced from multi-views, meanwhile, the structured sparse method also can be improved by the ℓ_0-norm regularization.

Essentially, our aim in this paper is to investigate the ignored mutual dependence among observations of the learned dictionaries from K-SVD, which would contain variation features about the uncontrolled variations for each subject. For such purpose, a new scheme called the KSVD-MSRC approach is proposed, which is different from the traditional SRC. The proposed approach will utilize the classification of framework via a revised K-SVD algorithm. Moreover, in order to achieve more robustness, we propose the mixed ℓ_0-norm regularization instead of the general ℓ_1-norm and ℓ_2-norm. Furthermore, based on recent advantages in the alternating direction method of multipliers (ADMM), the proposed KSVD-MSRC problem can be easily split into three subproblems. Compared with MADMM(manifold based on alternating direction method Of multipliers) [44] and SADMM(subspace based alternating

direction method Of multipliers) [43], the biggest difference is that our problem is nonconvex while other two problems are convex. In summary, our main contributions can be described as follows:

1. The dictionary is learned by employing the revised K-SVD model, which is employed to solve this optimization problem via the ℓ_0-norm optimization in each step. The reason is that the ℓ_0-norm is the essential measure of sparsity, and the ℓ_1-norm can only been seen as its best convex relaxation form.

2. The mixed ℓ_0-norm regularization model with ℓ_1 fidelity combines the advantages of SRC and JSRC. By virtue of the mixed ℓ_0-norm regularization, the proposed model is more robust when it deals with large pose variation images. To the best of our knowledge, this is the first time to integrate the mixed ℓ_0-norm regularization and ℓ_1 fidelity, which has demonstrated the superiority of performance.

3. An efficient algorithm using the alternating direction method of multipliers to optimize our KSVD-MSRC models is developed. Moreover, promising results from extensive comparison experiments on 3 benchmark databases validate the effectiveness of our method and demonstrate the great potential of KSVD-MSRC in pattern classification tasks.

The remainder of this paper is organized as follows. In section 2, we review the basic of dictionary learning. In section 3, the proposed algorithm KSVD-MSRC is described in details. In section 4, extensive recognition experiments are conducted on 3 databases: CMU-PIE, Multi-PIE and YaleB. Finally, we obtain a conclusion in section 5.

2 Related Work

2.1 Sparse Representation via ℓ_1-norm

In [39], the discriminative nature of sparse representation is exploited to perform classification. They assumed that any test sample can be represented as a linear combination of training samples from each class. This representation is naturally sparse, and involves only a small fraction of the overall dictionary. Based on this assumption, they introduced the sparse representation based classification (SRC). In sparse representation classification, given a set of gallery images $A = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^{m\times n}$, one seeks a sparse combination of these images to represent an unknown images Y. Such a sparse solution can be found by solving following problem in [39]:

$$\min_X \frac{1}{2}\|Y - AX\|_2^2 + \lambda\|x\|_1,$$

where λ is a tuning parameter, and $\|X\|_1$ is defined as the sum of absolute values of all entries. This convex optimization problem can be solved efficiently with many algorithms developed specifically for Compressed Sensing (CS). Then SRC combines this sparse representation with nearest subspace classification. In other words, it computes the class-specific residual vector and the target Y is classified according to the minimum ℓ_2-norm of the residual vectors:

$$\begin{aligned} r_k &= Y - A_k X_k, \\ class(Y) &= \arg\min_k \|r_k\|_2^2, \end{aligned}$$

where A_k is the sub-matrix of A that corresponds to all gallery images in class k, and X_k is the sub-vector of X with the corresponding sparse coefficients.

2.2 IIHT Algorithm via ℓ_0-norm

To make the ℓ_0-norm algorithm converge to a point satisfying certain optimality conditions (stationarity), a proper selection of step size is necessary at each iteration. An adaptive step size rule based on the restricted isometry property [3] to ensure a sufficient decrease in the objective per iteration for the classical linear compressed sensing is introduced. Recently, for the nonlinear sparse optimization problem, a non-monotone line search to ensure its convergence is proposed [24]. The remaining part of IIHT just follows the original IHT, which is described as follows,

$$\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_0,$$

where $\|x\|_0$ is defined as the number of nonzero entries. We emphasize that the major computation is very easy to obtain via solving,

$$arg\min\{\|y - x\|_2^2, \text{ s.t. } y \in \mathbb{R}^n\}, \tag{2.1}$$

By considering the restricted strong convexity and smoothness of $f(x)$, the convergence of IIHT can be established. Suppose that the function $f(x)$ has the Lipschitzian gradient with L-stationary point L_s [2]. For any $x, y \in \mathbb{R}^n$ satisfying $\|\Gamma_{xy}\|_0 \leq s$, one has

$$\|(\nabla(f(y) - \nabla f(x))\Gamma_{xy}\|_2^2 \leq L_s\|y - x\|_2^2.$$

where $\nabla f(x)$ is the gradient of $f(x)$ on $\mathbb{R}^n$. Let the sequence $\{x^k\}$ be generated by IIHT. It can be shown that the linear convergence rate both in terms of functional value sequence $\{f(x^k)\}$ and the sequence itself $\{x^k\}$. That is, for any sufficiently large k, it holds,

$$\|x^{k+1} - x^*\|_2^2 \leq \rho\|x^k - x^*\|_2^2,$$

where ρ is a constant, x is any s-sparse and obeys $y = Ax$.

We note that the above fast convergence is significant in the analysis of ℓ_0-norm algorithm. The numerical experiments demonstrate IIHT can work as well as original ℓ_0-norm in reconstruction ability and accelerate the speed of spare coding, according to the DL algorithm process.

2.3 Synthesis Dictionary Learning

The main idea of SDL is to approximately reconstruct the original samples by the combination of dictionary atoms with respective weight factors. The weight factors are stored in the form of coefficients. Let $Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{m\times n}$ be the original data matrix, each column of which represents the m-dimensional feature of one sample. And let $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{c\times n}$ be the coding coefficients of Y over a learnt dictionary. The basic formula of SDL is presented as the following:

$$\begin{aligned} &\min_{D,X} \|Y - DX\|_F^2 \\ &\text{s.t. } D \in \mathcal{A}, \ \|x_i\|_0 \leq T_0, \ \forall i, \end{aligned}$$

where $D \in \mathbb{R}^{m\times c}$ is the synthesis dictionary, $\mathcal{A}$ denotes a set of constraints on D to make the solution non-trivial, and T_0 is the sparsity level. Euclidean distance is always an effective tool for distance metric. In the above formula, minimizing the square of Frobenius norm means minimizing the sum of all the distances between original data and its approximate representation, i.e. minimizing the residual.

The column dimension represents the poses/images in the gallery as grouped by subjects, while the row dimension represents different images in the unknown set Y. Each column denotes a sparse representation vector and each square denotes a coefficient. (a) Independent sparsity (as in SRC): all coefficients are selected independently based on ℓ_1 regularization; (b) IIHT based on ℓ_0-norm: nonconvex penalty functions perform better than the ℓ_1-norm in terms of estimation accuracy and consistency; (c) Dictionary learning: Update the D and X used ℓ_0-norm algorithm of OMP(Orthogonal Matching Pursuit); As mention above, a new mixed sparsity should be the highest priority requirement to balance both ℓ_0 and ℓ_1 to adaptively select the suitable class-level and atom-level sparsity.

3 The Proposed Model

The main idea of proposed method is to focus on two parts. Firstly, it is to learn the discrimination dictionary on which we can gain effective coefficients of each sample. Secondly, the mixed ℓ_0-norm sparse representation classification method to overcome the issues with JDSRC is presented.

3.1 Dictionary Learning with Synthesis K-SVD

The dictionary learning methods aim to learn a good dictionary from original training samples such that it can properly represent the original samples in feature extraction process. But due to original K-SVD embedding with OMP method has the unsupervised nature characteristic, it suffers from low classification accuracy and slow training speed. Consequently, a new proposed method embedding with IIHT algorithm is to learn an optimized dictionary with which we can gain effective coefficients of each sample with fast training speed. However, section 2.3 only focuses on representative power without considering discrimination ability of dictionary. For performing pattern classification tasks better, we expect that images from the same class have similar representations. This problem can turn to handle the problem of learning a dictionary which impels coefficients obtained from intra-class images mostly similar and close but maximizes the dissimilarity of inter-class coefficients. Mathematically, in the original K-SVD formulation, Y can be any face data for one subject and D is the learned dictionary representing this subject.

Assume there are totally P subjects and n images per person in the training set, we can write Y as $Y = [Y_1, Y_2, \ldots, Y_p] \in \mathbb{R}^{d\times N}$. Hence, Y_i can be rewritten as $y_i = [y_{i1}, y_{i2}, \ldots, y_{in}] \in \mathbb{R}^{d\times n}$, where y_{ij} is a d-dimensional vector of cropped face image. In order to extract more discriminative and robust information from this training set, K-SVD aims to simultaneously learn a discriminative structured dictionary for all individual image sets for each subject, on which each image is encoded by a discriminative coefficient. To achieve this, we formulate the following optimization problem:

$$\min_{D_i, X_i} \|Y_i - D_i X_i\|_F^2 + \lambda \|X_i\|_0,$$

where $Y_i \in \mathbb{R}^{d\times n}$ is a group of samples from the training subject, and each column of Y_i denotes a face vector, X_i is the coefficient vector of Y_i, which is the sparse representation of

training samples in Y. In this paper, we choose Pan's [22] method to use the IIHT algorithm for faster convergence. Here, $D = [D_1, D_2, \ldots, D_p]$ is a structured dictionary learned from the K-SVD, X_i is simultaneously penalized by the same parameter λ in this paper. In detail, this term enforces data from each class have their own target different expressions after transformation. That is to say, the coefficients X_i for one subject are similar with ones from identical category while distinct with ones from other classes. An appropriate value of parameter can make our model be balanced with discriminative power in next subsection.

In general, the difference of sparse codes for two face images should be minimized if they are from the same class as they should look similar, the difference of sparse codes for two face images should be maximized if they are from different classes as they should be different. In this case, the images of per subject are separately trained by the K-SVD to obtain different dictionaries. It is reasonable that combining all dictionaries forms an integrated dictionary which has the standard class labels corresponding to the image labels before trained. So, after the group sparse coding and dictionary updating, we can archive a collection of X and D with class labels, where $X = [X_1, X_2, \ldots, X_n]$ and $D = [D_1, D_2, \ldots, D_p]$. In detail, the update of k-th column of each subject is done by rewriting the penalty term as

$$\begin{aligned}\|Y_i - D_i X_i\|_F^2 &= \|(Y_i - \sum_{j \neq k} d_j x_T^j) - d_k x_T^k\|_F^2 \\ &= \|E_k - d_k x_T^k\|_F^2\end{aligned}$$

where X_T^k denotes the k-th row of X_i. After this step, we preserve the matrix Ω_k with a size $n \times |\{i|1 \leq i \leq K, x_T^k(i) \neq 0 \bigcap x_T^k(i) \leq T\}|$ according to the sparsity [1]. The same effect happens with $E_k^R = E_k \Omega_k$, implying a selection of error columns that correspond to examples that use the atom d_k. So the minimization problem as mentioned before becomes $\|E_k^R - d_k x_T^k \Omega_k\|_F^2$, which can be solved into $U \Delta V^T$ by using the SVD. Then we can obtain that, $d_k = U(:, 1)$ and the coefficient vector $x_k = x_T^k \Omega_k$ is actually the first column of $V \times \Delta(1, 1)$. Once the entire dictionary is updated, the sparse coding process will be invoked again and then we can update D consequently. The details of the solution of K-SVD is described as below:

In summary, the update scheme for the K-SVD can be rapidly implemented in each iteration, and the discriminate dictionary is achieved by minimizing the following subsections.

3.2 The Proposed KSVD-MSRC for Robust Sparse Representation

In this section, we aim to modify the SRC algorithm based on the revised K-SVD algorithm. In fact, by combining the learned dictionaries from the proposed K-SVD for each subject, the joint sparse representation classification (JSRC) [27] can be modified to exploit the shared information from all the samples. To describe our model,we first extend the original SRC to the following robust and stable formulation for sparse representation:

$$\min_X \ \frac{1}{2}\|D_Y - D_A X\|_1 + \lambda \|X\|_0,$$

where $D_Y \in \mathbb{R}^{d \times N}$ represents the probe dictionary matrix that contains all the samples in probe set, and $D_A \in \mathbb{R}^{d \times m}$ is the dictionary matrix that learned from the gallery set. Next, we introduce information sharing between different face views. Recall that each column of the coefficient matrix $X \in \mathbb{R}^{m \times N}$ represents one view of a subject, and each row represents the weights of the corresponding gallery images in all views of the same subject. The same

Algorithm 1: Revised KSVD

1: **Task** Find the best dictionary to represent the data samples Y.
2: **Initialize** $D_0 \in \mathbb{R}^{d\times N}$, set $J = 1$
3: **repeat**
4: **Sparse Coding Step:** Using the IIHT technique in [23] to compute the representation vectors $X = [X_0, X_1, \ldots, X_N]$.
5: **Codebook Update Step:** For each column $k = 1, 2, \ldots, n$ in D^{J-1},update it by

- Calculate the error matrix $E_k = Y - \sum_{j\neq k}^{n} d_j x_k^j$,
  Restrict E_k by choosing only the columns corresponding to $|\{i|1 \leq i \leq K, x_T^k(i) \neq 0 \bigcap x_T^k(i) \leq T\}|$, and obtain E_k^R
- Using SVD decomposition $E_k^R = U \triangle V^T$
- Update the dictionary column $d_k = U(:, 1)$
- Update the cloned vector $x_k = V(:, 1) * \triangle(1, 1)$

6: **until** Set $J = J + 1$.

hypotheses are applied with JSRC, the shared information appears in each face image for one subject onto the previous formulation. The $\ell_{2,0}$-norm is used on the coefficients matrix X to exploit the shared information. To the best of our knowledge, we are the first to model the mixed $\ell_{2,0}$-norm problem JSRC framework with nonconvex regularization. In comparison to the convex $\ell_{2,1}$-norm, this nonconvex $\ell_{2,0}$-norm enables us to extract a higher performance. In this paper, we propose the following optimization model:

$$\min_X \ \|D_Y - D_A X\|_1 + \lambda \|X\|_*, \tag{3.1}$$

where the mixed ℓ_0-norm is defined as

$$\|X\|_* = \gamma \|X\|_0 + (1-\gamma)\|X\|_{2,0}.$$

Here, $\|X\|_{2,0}$ is the cardinality of the ℓ_2-norm of all rows of a matrix. The parameter γ controls the trade-off between ℓ_0-norm and $\ell_{2,0}$-norm.

To make use of the structure of (3.1), we then formulate (3.1) as an equivalent constrained problem

$$\begin{aligned} &\min_{X,V,Z,T} \|V\|_1 + \gamma\lambda\|Z\|_0 + (1-\gamma)\lambda\|T\|_{2,0} \\ &\text{s.t.} \quad D_A X - D_Y - V = 0,\ X - Z = 0,\ X - T = 0. \end{aligned} \tag{3.2}$$

The associated augmented Lagrangian function is defined as follows

$$\begin{aligned} &\mathcal{L}(X, V, Z, T, W_1, W_2, W_3) \\ &= \|V\|_1 + \gamma\lambda\|Z\|_0 + (1-\gamma)\lambda\|T\|_{2,0} \\ &-tr[W_1^T(D_A X - D_Y - V)] + \frac{\sigma_1}{2}\|D_A X - D_Y - V\|_F^2 \\ &-tr[W_2^T(X - Z)] + \frac{\sigma_2}{2}\|X - Z\|_F^2 \end{aligned}$$

$$-tr[W_3^T(X-T)] + \frac{\sigma_3}{2}\|X-T\|_F^2,$$

which can be simplified to

$$\begin{aligned}
&\mathcal{L}(X,V,Z,T,W_1,W_2,W_3)\\
&= \|V\|_1 + \gamma\lambda\|Z\|_\Phi + (1-\gamma)\lambda\|T\|_{2,\Phi}\\
&+\frac{\sigma_1}{2}\|D_AX - D_Y - V - W_1/\sigma_1\|_F^2\\
&+\frac{\sigma_2}{2}\|X - Z - W_2/\sigma_2\|_F^2\\
&+\frac{\sigma_3}{2}\|X - T - W_3/\sigma_3\|_F^2 + C,
\end{aligned}$$

where W_1, W_2, W_3 are the Lagrange multipliers, $\sigma_1, \sigma_2, \sigma_3$ are positive penalty parameters, $tr(\cdot)$ is the trace of matrix, and C is a constant. It is difficult to simultaneously optimize all these variables. We therefore approximately solve this optimization problem by alternatively minimizing one variable with the others fixed. Under the framework of multi-block ADMM, the optimization problem of $\mathcal{L}$ with respect to each variable can be solved by the following steps. Now, we will show that each step of the ADMM either has a closed-form solution or can be solved by a fast solver.

Step 1: For variable V, the optimization subproblem of $\mathcal{L}$ with respect to V is equivalent to

$$V^{k+1} = \arg\min_V \|V\|_1 + \frac{\sigma_1}{2}\|D_AX^k - D_Y - V - W_1^k/\sigma_1\|_F^2.$$

From the following lemma, the solution can be given by the following soft-shrinkage operator

$$V^{k+1} = \text{shrink}_{2,1}(D_AX^k - D_Y - W_1^k/\sigma_1, 1/\sigma_1). \tag{3.3}$$

Lemma 3.1 The minimization problem

$$\min_x \ \lambda\|x\|_1 + \frac{1}{2}\|x-t\|_2^2$$

with $\lambda > 0$ and $t \in \mathbb{R}$ has a closed-form solution, and it is given by the following soft-shrinkage operator

$$x^* = \text{shrink}_{2,1}(t,\lambda) := \text{sign}(t) \circ \max\{|t| - \lambda, 0\},$$

where $\circ$ and sign represent, respectively, the point-wise product and the signum function, and all operations are done componentwise. See [9] and [11] for more details.

Step 2: For the variable Z, optimizing $\mathcal{L}$ with respect to Z can be simplified to

$$Z^{k+1} = \arg\min_Z \gamma\lambda\|Z\|_0 + \frac{\sigma_2}{2}\|X^k - Z - W_2^k/\sigma_2\|_F^2.$$

According to the lemma below, the solution is

$$Z^{k+1} = \text{shrink}_{2,0}(X^k - W_2^k/\sigma_2, 2\gamma\lambda/\sigma_2). \tag{3.4}$$

Lemma 3.2 The minimization problem

$$\min_x \ \lambda\|x\|_0 + \|x-t\|_2^2$$

with $\lambda > 0$ and $t \in \mathbb{R}$ has a closed-form solution which is given by the following hard-shrinkage operator

$$x^* = \text{shrink}_{2,0}(t, \lambda) := \begin{cases} t & \text{if } |t| > \lambda^{0.5}, \\ 0 & \text{if } |t| \leq \lambda^{0.5}. \end{cases}$$

Step 3: For variable T, the subproblem of $\mathcal{L}$ with respect to T can be transformed as

$$T^{k+1} = \arg\min_T \ (1-\gamma)\lambda\|T\|_{2,0} + \frac{\sigma_3}{2}\|X^k - T - W_3^k/\sigma_3\|_F^2.$$

Notice that the optimizations of each row are independent of each other. Therefore, it can be decomposed into

$$t_i^{k+1} = \arg\min_{t_i} \ (1-\gamma)\lambda\|t_i\|_2 + \frac{\sigma_3}{2}\|x_i^k - t_i - (w_3^k)_i/\sigma_3\|_2^2,$$

where $t_i, x_i^k, (w_3^k)_i$ are the i-th row vectors of T, X^k, W_3^k, respectively. Based on the following lemma, it can be solved by

$$t_i^{k+1} = \text{shrink}_{2,2}(x_i^k - (w_3^k)_i/\sigma_3, (1-\gamma)\lambda/\sigma_3).$$

Lemma 3.3 The minimization problem

$$\min_x \ \lambda\|x\|_2 + \frac{1}{2}\|x - t\|_2^2$$

with $\lambda > 0$ and $t \in \mathbb{R}$ has a closed-form solution, and it is given by

$$x^* = \text{shrink}_{2,2}(t, \lambda) := \frac{t}{\|t\|_2} \circ \max\{\|t\|_2 - \lambda, 0\}.$$

Hence, the solution of this subproblem is

$$T^{k+1} = \text{shrink}_{2,0}(T^{k+1}, 2(1-\gamma)\lambda/\sigma_3). \tag{3.5}$$

Step 4: For variable X, the subproblem of $\mathcal{L}$ with respect to X is equivalent to the linear system

$$\begin{aligned}(\sigma_1 D_A^T D_A + \sigma_2 I + \sigma_3 I)X = \sigma_1 D_A^T (D_Y + V^k + W_1^k\ /\sigma_1) \\ + \sigma_2(Z^k + W_2^k\ /\sigma_2) + \sigma_3(T^k + W_3^k\ /\sigma_3),\end{aligned}$$

which yields,

$$\begin{aligned}X^{k+1} = (\sigma_1\sigma_1 D_A^T D_A + \sigma_2 I + \sigma_3 I)^{-1} \\ (\sigma_1 D_A^T (D_Y + V^k + W_1^k\ /\sigma_1) + \sigma_2(Z^k + W_2^k\ /\sigma_2) \\ + \sigma_3(T^k + W_3^k\ /\sigma_3)).\end{aligned} \tag{3.6}$$

In practice, it can be efficiently solved by applying Cholesky decomposition.

Step 5: For dual variables W_1, W_2, W_3, according to the ADMM, the multipliers associated with $\mathcal{L}$ are updated by the following formulas

$$W_1^{k+1} = W_1^k - \sigma_1(D_A X^{k+1} - D_Y - V^{k+1}), \tag{3.7}$$

$$W_2^{k+1} = W_2^k - \sigma_2(X^{k+1} - Z^{k+1}),$$
$$W_3^{k+1} = W_3^k - \sigma_3(X^{k+1} - T^{k+1}).$$

By summarizing the above steps, one can obtain the solutions by Algorithm 2, and the convergence is given as follows.

Theorem Suppose that $\{(V^k, Z^k, T^k, X_1^k, W_1^k, W_2^k, W_3^k)\}$ is a sequence generated by Algorithm 2, then any cluster point $(V^*, Z^*, T^*, X_*, W_1^*, W_2^*, W_3^*)$ of the sequence is a stationary point of (3.2).

The proof follows a similar line of arguments as in [14] and [42], thus we omit the detail for succinctness.

Algorithm 2: - The KSVD-MSRC solution based on the ADMM

1: **input**
Probe face dictionary $D_Y \in \mathbb{R}^{d\times N}$, galley dictionary set $D_A \in \mathbb{R}^{d\times m}$, regularization parameters λ and γ, penalty parameters σ_1, σ_2 and σ_3 .
2: **Initialize**
$V \in \mathbb{R}^{d\times N}$, $Z \in \mathbb{R}^{m\times N}$, $T \in \mathbb{R}^{m\times N}$, $W_1 \in \mathbb{R}^{d\times N}$, $W_2 \in \mathbb{R}^{m\times N}$, $W_3 \in \mathbb{R}^{m\times N}$.
3: **repeat**
4: **While not converge do**
5: **1. Update V^{K+1} according to (3.3)**
6: **2. Update Z^{K+1} according to (3.4)**
7: **3. Update T^{K+1} according to (3.5)**
8: **4. Update X^{K+1} according to (3.6)**
9: **5. Update W_i^{K+1} according to (3.7)**
10: **until** Some stopping criterion

3.3 Recognition and Classification

The fitness for class k is represented by the residual matrix

$$E_k = D_Y - D_A^k X^k,$$
$$Identify(Y) = \arg\min_k \|E_k\|_F^2,$$

where D_A^k is the k-th column of the training set D_A, and X^k is the coefficient matrix corresponding to each subject. In summary, the proposed method is based on a joint sparse representation with the revised K-SVD on ℓ_0-norm and $\ell_{2,0}$-norm optimization. In fact, the mixed norm optimization for sparse representation in motivated by the following observations: 1) The input samples are assumed to lie on dictionary learning from the revised K-SVD learning method for face images; 2) The optimized variables can be transferred to a mixed norm optimization if the variable and input data is linearly related as demonstrated in [47]. 3) The treatments on optimization to the K-SVD and KSVD-MSRC are different.

4 Numerical Experiments

In this section, we can apply our algorithms to solve the mixed norm model and compare them with the other methods. All the experiments are performed using MATLAB (R2017a)

on a desktop computer with an Intel Xeon E5 4.0 GHz Dual Core with 32 GB RAM to show effectiveness of the proposed approach. Three popular face databases, including the CMU-PIE, YaleB and Multi-PIE, which are shown in the Fig.1 are used. In order to show the performance of our proposed method, three parts of experiments are conducted, including: 1) Parameters set-up, 2) The proposed K-SVD and convergence analysis for the KSVD-MSRC, 3) Performance analysis for face recognition.

Figure 1: From top to bottom are CMU-PIE, YaleB, Multi-PIE databases which is described in our experiment.

4.1 Parameters Set-up

For notational convenience, the novel method is denoted as KSVD-MSRC. For the parameters of K-SVD model, here the dictionary length is assumed to be the length of the input samples. In spare representation, we follow the standard cross validation procedure in machine learning to select the regularisation parameter λ for SRC and JSRC. This is achieved by further dividing the training set into a smaller training set and a validation set. For the proposed KSVD-MSRC method, we also follow the same procedure, wherein all the training, validation, and test sets are exactly the same as those used for SRC and JSRC. The only minor difference is that KSVD-MSRC has both the regularisation parameter λ and the mixed norm parameter γ ($\gamma \in [0, 1]$). All initial values $(W_1^0, W_2^0, W_3^0) = (0, 0, 0, 0)$. As is well know, the accuracy of the solution depends on the value of the regularization parameters λ and γ. Thus, we will try values from the set $\{0.5, 0.1, 0.05, 0.01, 0.005,$
$0.001\}$ and select the one that give the highest recognition rate. Regarding the penalty parameters $\sigma_1, \sigma_2, \sigma_3$, we also try some values and choose a satisfactory performance values. This means the computation slightly increases because the search is done on two dimensions. From the interpretation of the role of in controlling the pose variation and our intensive numerical studies, we suggest that the computational increase in cross validation might be

reduced by a preliminary estimation of the pose variations.

4.2 The proposed K-SVD and Convergence Analysis for the KSVD-MSRC

In this section, the performance of proposed K-SVD and convergence of KSVD-MSRC will be evaluated. As motivated by discrimination representation, novel method should have the effect of compacting dictionaries and preserving feature similarity. To evaluate the compactness of the learned dictionary from the proposed K-SVD, the embedded IIIHT with $\lambda = 0.0005$, $\sigma = 0.4$ is employed to calculate the normalized orthogonality in face images. Since the ℓ_0-norm problem is a convex problem which can be iteratively solved with high efficiency, so the reconstruction error it is an important indicator in this experiment. In Fig.2, the different number of coefficients and iterations are considered. It show the reconstruction error of the YaleB database based revised K-SVD algorithm with the sparsity level 5 to 35, and it is clear that with the increasing sparsity, the error is decreased, and we also can see that all the cases reach the reconstruction error threshold after 12 iterations. These show that revised K-SVD algorithm always converges rapidly regardless of the dictionary size.

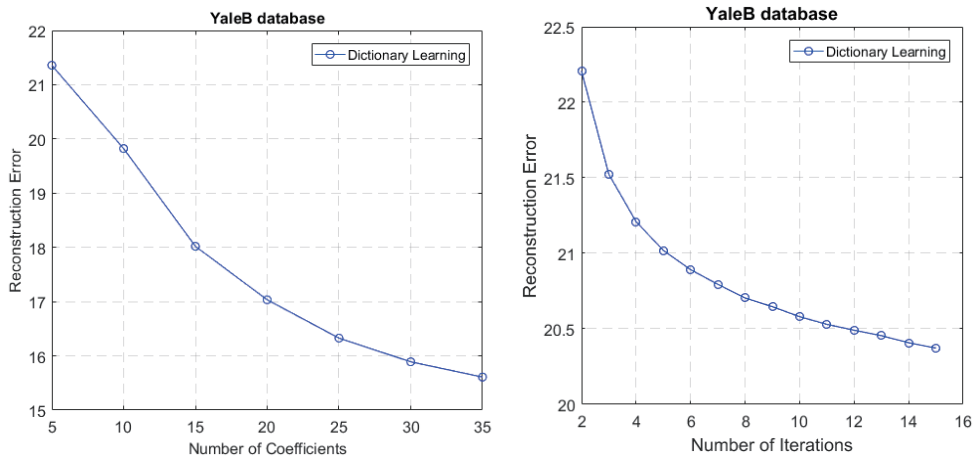


Figure 2: An example of revised K-SVD process on the YaleB face database. (a) the reconstruction error via same sparsity. (b) the reconstruction error via same iteration.

In addition, the RMSE(root mean square error) and time of recognition process is also considered for face recognition performance. In the sequent section, the computation time of K-SVD is calculated when it has 40 images per person on the CMU-PIE database. For a fair and thorough comparisons, 100 subjects are recorded on each database respectively for the computational time. Iteration number is set as 40. All images are resized to $32 \times 32 = 1024$, and the training samples are randomly chosen from the database, the final results are shown in Fig. 3, with the increase of the dictionary size, the RMSE of proposed K-SVD will be gradually stable and smaller than the original K-SVD, and the computational time of proposed K-SVD increases slower than the K-SVD when the sample number is sufficiently large.

The above results show that the proposed K-SVD can achieve the better performance with compact characterization. Its effectiveness on face recognition based on the proposed KSVD-MSRC will be conducted based on the ADMM framework. When we setup the proper stopping criterions, the proposed ADMM-based method can converge to modest accuracy

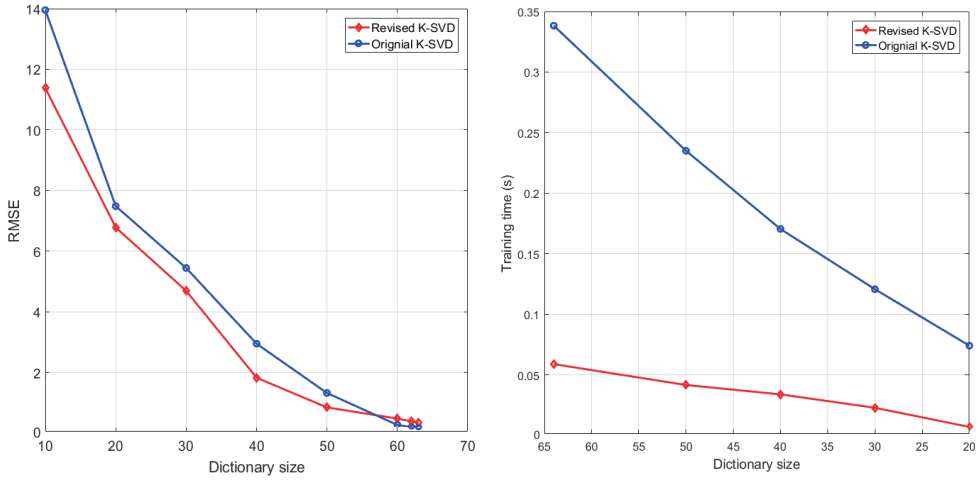


Figure 3: Illustration of the RMSE and time cost via orignial K-SVD and revised K-SVD on CMU-PIE database.

within a few tens of iterations, this behavior makes our method deal with large-scale problem in a short time. It can be seen from the Fig.4, with the increasing iteration, the objective function is decreasing and stable after 30 iterations.

4.3 Face Recognition with different Number of Views

In order to show the performance of related methods under different number of views, we follow the experiment settings in [45] for CMU-PIE data set. The views subset $[0°, \pm 22.5°, \pm 45°, \pm 67.5°, \pm 90°]$ are chosen for the training set. Only one face image is selected for each subject with each pose in the training, and only one image for each pose. Since we randomly select from all 13 poses, the selected pose may not exist in the training set. For the YaleB data set, the same setting is also applied, which will makes our experiments more realistic and challenging.

In Fig.5, the classification accuracy of the proposed method both on CMU-PIE and YaleB are compared with others. As it can be seen, traditional subspace methods cannot reach satisfactory classification rates, but all SRC based methods can work well in multiple-views scenario. If there is just one testing image, none of methods can perform well. We note that all methods perform better in YaleB than CMU-PIE when there is only one training image. The reason for this is that the YaleB only has 10 subjects, which is much less than CMU-PIE. When more views are added in, the performance of SRC based methods is increased. Especially, our proposed KSVD-MSRC reached a satisfactory rate, and it achieved 97.82% when we have 7 views in the test set. Clearly, this outperforms the closest competitor JSRC by about 1% for CMU-PIE. It can be further improved by adding more number of views for YaleB. Furthermore, it can be noted that both KSVD-MSRC and MSRC have a similar recognition rate on CMU-PIE. When the number of views increases, MSRC cannot achieve the same performance as KSVD-MSRC.

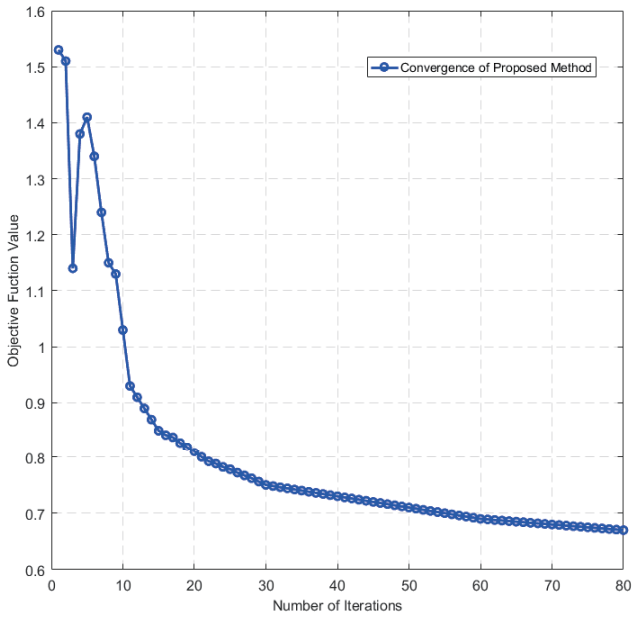


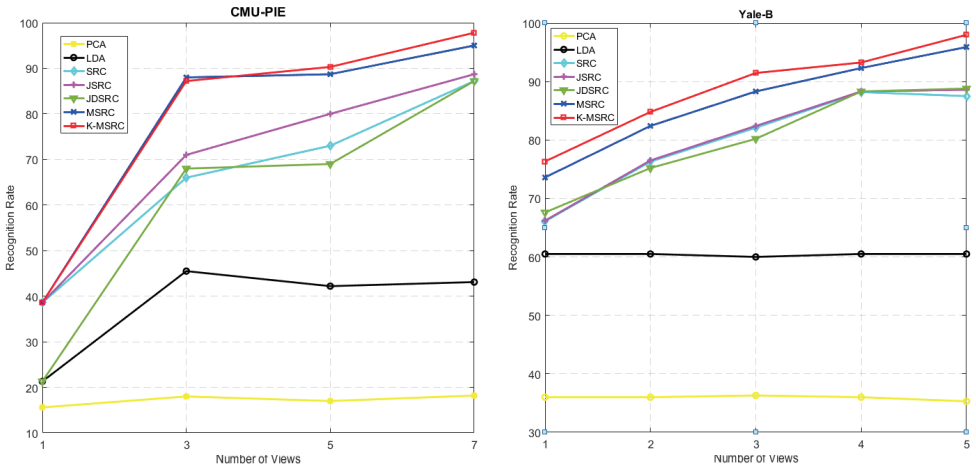
Figure 4: Convergence analysis of the proposed method. The objective function is converged after 30 iterations.

4.4 Face Recognition under Different Dimensions

In this subsection, the performance of the face recognition based on different feature dimensions is investigated. Prior to the experiment, the original image of CMU-PIE is reduced to the $d = [32, 64, 128, 256]$, which is effective for SRC introduced in [39]. Following the previous experiments of [45], the same training set are employed and randomly 5 views are chosen for the testing set. For YaleB data set, the dimension is reduced to $d = [8, 16, 32, 64, 128]$ since the pose variation is not as large as CMU-PIE,. It can be seen the comparison results from the Fig.6, which shows that the proposed methods have obtained much higher performance than the other methods under the same dimensions for CMU-PIE and YaleB databases. MSRC is a little less than KSVD-MSRC method. Though less than KSVD-MSRC, both JSRC and SRC have competitive accuracy. The performance of all these methods are superior to JDSRC in most image dimensions. When the data dimension ≥ 64 in CMU-PIE and 32 in YaleB, the performance becomes saturated in all these methods. However, the recognition rate of JDSRC is reduced after $d = 128$ in CMU-PIE, which is caused by the low accuracy of its greedy algorithm. Considering the case of methods, when the dimension ≥ 64 for CMU-PIE or 32 for YaleB, KSVD-MSRC is not sensitive to feature dimensions. That means it is a good choice with 64 feature dimension for our method. Additionally, the satisfactory performance with much lower computational complexity also can be achieved.

4.5 Face Recognition under Different Views

Under large view variations in multiple face recognition, the performance of the proposed method is investigated with other methods. Since CMU-PIE data set has more than 9 poses, which is sufficient to perform this experiment, so only CMU-PIE is employed in this subsection. We followed the same setting as follows in [45]. Face images of all 13 poses are



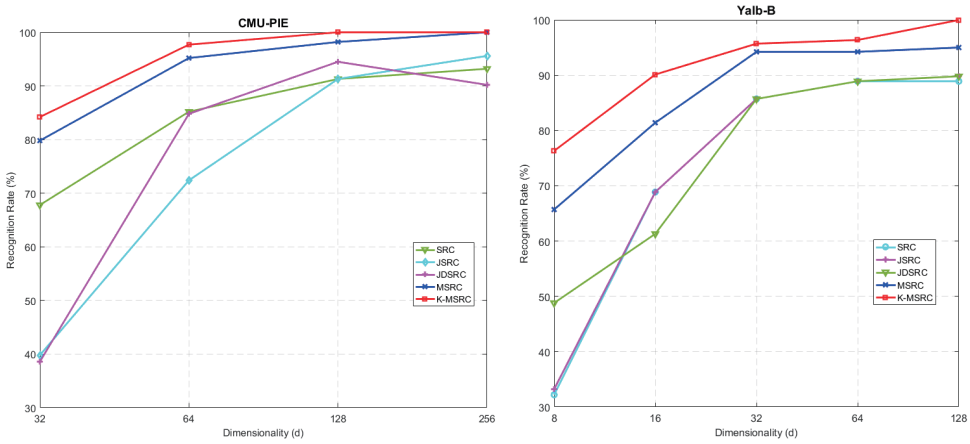
(a) (b)

Figure 5: Face recognition under different number of views with dimension d=64 for CMU-PIE and YaleB.

Methods	22.5°	45°	67.5°	90°
PCA	24.43	24.68	24.38	23.43
LDA	59.75	60.00	58.11	54.13
SRC	82.24	80.30	80.15	69.25
JSRC	84.48	84.63	83.28	74.18
JDSRC	87.16	84.63	87.16	78.21
MSRC [47]	95.52	94.33	93.73	88.96
KSVD-MSRC	**97.54**	**95.73**	**95.13**	**89.67**

Table 1: Face recognition against large view variations

employed for the training set, but only one image is randomly selected for each pose of per subject. We then create 4 different view groups: $[0°, \pm 22.5°]$, $[0°, \pm 45°]$, $[0°, \pm 67.5°]$, and $[0°, \pm 90°]$. There are 3 images (one image for each view) in each group. The testing sets are generated by randomly selecting from these 4 view groups. As can be seen in Table 1, traditional subspace methods perform poorly, which demonstrate consistency with previous experimental results. However, all SRC-based methods obtain satisfactory performance. It can be observed that when the view difference increases, the performance of SRC decreases. JSRC performs slightly better than SRC across all view variations. Also, JDSRC outperforms both JSRC and SRC. Since JDSRC uses dynamic selected atoms, it would not select the same set of atoms for all views as JSRC. This makes JDSRC more suitable to multiple-views scenarios. Overall, our proposed method reaches the highest recognition rates under each testing view group. It achieves 97.54% for $[0°, \pm 22.5°]$ group with 2% improvement compared to its best competitor.



(a) (b)

Figure 6: Face recognition under different dimensional with same views for CMU-PIE and YaleB.

4.6 Face Recognition with a Large Number of Subjects

In order to detect the compared methods with a large number of subjects, the Multi-PIE data set is employed to perform two sets of experiments. To make our experiments more realistic and challenging, all images from the 337 subjects with 102 females and 235 males are mixed. Images in each training set are selected based on following views $[0°, \pm 30°, \pm 60°, \pm 90°]$. Three face images are selected for each subject with each view in the training. For testing, views are randomly selected from all views for each subject (5 or 7). We also employ three face images for each view in testing set. Four pairs of training and testing sets are created from randomly selected subjects from all 337 subjects and ten sets are created. These pairs of data set correspond to 64 (34 females and 34 males), 136 (68 females and 68 males), 204 (102 females and 102 males), 272 (all random) subjects. All subjects were randomly chosen from 337 subjects. For obtaining the fair experiment results, 10 different random data sets are generated, and at least 65 subjects are left for random selection purpose.

It can be seen from the Table 2, when the number of subjects increases, the performance of SRC, JSRC, JDSRC, MSRC and KSVD-MSRC deteriorates significantly. However, the proposed KSVD-MSRC is more robust against this large number of subjects. It also shown that all methods are benefit from decreasing number of subjects. Their performance are observed to improve overall, there is less sharp drop in recognition accuracy when number of subjects increases. Due to the flexible atom selection, both KSVD-MSRC and MSRC outperform SRC and JSRC. However, the lack of guaranteed convergence of JDSRC makes it hard to find a robust and accurate solution. In general, the proposed method achieves a robust performance against different scales of data sets because it has an advantage of a dynamic atom selection and fast convergence.

5 Conclusions

In this paper, by introducing the definition of mixed ℓ_0-norm, a novel framework of KSVD-MSRC is proposed to establish the optimization problem. Firstly, we propose an dictionary learning approach based on K-SVD embedding with IIHT algorithm for faster convergence

Methods	68	136	204	272
SRC	86.74	82.30	78.55	77.47
JSRC	85.24	79.30	76.15	73.56
JDSRC	89.16	85.98	82.26	80.21
MSRC [47]	92.52	93.12	90.73	90.73
KSVD-MSRC	**93.20**	**95.63**	**93.24**	**92.60**

Table 2: Face recognition with different number of subjects for Multi-PIE database

to construct the relationship between training and testing set. Secondly, mixed norm penalty of ℓ_0-norm and $\ell_{2,0}$-norm is embedded into sparse representation classification to regulate the sparsity of inter-class and inner-class level, which has demonstrated to outperform other state-of-the-art methods. Furthermore, KSVD-MSRC is solved on the powerful ADMM framework in deriving the numerical algorithm, which further improve the performance and computation. Experimental results on three different databases demonstrate superior performance of the proposed methods under a different number of views, various dimensionality, view differences, and scalability.

References

[1] M. Aharon, M. Elad, Michael and A. Bruckstein, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, *IEEE Transactions on signal processing* 54 (2016) 4311–4322.

[2] A. Beck and Y. Eldar, Sparsity constrained nonlinear optimization: Optimality conditions and algorithms, *SIAM J. Optim.* 23 (2013)1480–1509.

[3] E.J. Candes and T. Tao, Decoding by linear programming, *IEEE Transactions on Information Theory* 51 (2005) 4203–4215.

[4] H. Cevikalp and B. Triggs, Face recognition based on image sets, in: *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on, 2010, pp. 2567–2573.

[5] Y.-C. Chen, V.M. Patel, P.J. Phillips and R. Chellappa, Dictionary-based face recognition from video, in *European Conference on Computer Vision*, Springer, 2012, pp. 766–779.

[6] W. Deng, J. Hu and J. Guo, Extended src: Undersampled face recognition via intraclass variant dictionary, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (2012) 1864–1870.

[7] W. Deng, J. Hu and J. Guo, In defense of sparsity based face recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 399–406.

[8] R.-X. Ding, D.K. Du, Z.-H. Huang, Z.-M. Li and K. Shang, Variational feature representation-based classification for face recognition with single sample per person, *J. Visual Comm. Image Represen.* 30 (2015) 35–45.

[9] D.L. Donoho, De-noising by soft-thresholding, *IEEE Transactionson Information Theory* 41 (1995) 613–627.

[10] D. Donoho and Y. Tsaig, Fast solution of l1-norm minimization problems when the solution may be sparse, 2006, Preprint, vol.1, no.2.

[11] M. Elad, Why simple shrinkage is still relevant for redundant representations?, *IEEE Transactions on Information Theory* 52 (2006) 5559–5569.

[12] S. Gao, I.W.-H. Tsang, and L.-T. Chia, Kernel sparse representation for image classification and face recognition, in: *European Conference on Computer Vision*, Springer, 2010, pp. 1–14.

[13] A. Hadid and M. Pietikainen, From still image to video-based face recognition: an experiment alanalysis, in: *Automatic Face and Gesture Recognition, Proceedings. Sixth IEEE International Conferenceon*, 2004, pp. 813–818.

[14] M. Hong, Z.-Q. Luo, and M. Razaviyayn, Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems, *SIAM J. Optim.* 26 (2016) 337–364.

[15] Y. Hu, A.S.Mian and R. Owens, Sparse approximate dnearest points for imageset classification, in: *Computer vision and pattern recognition (CVPR)*, 2011, pp. 121–128.

[16] Y. Hu, A.S. Mian and R. Owens, Face recognition using sparse approximated nearest points between image sets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (2012) 1992–2004.

[17] G.B. Huang, M. Ramesh, T. Berg and E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, in: Technical Report, University of Massachusetts, 2007, pp. 7–49.

[18] L. Huang, J. Lu, and Y.-P. Tan, Co-learned multi-view spectral clustering for face recognition based on image sets, *IEEE Signal Processing Letters* 21 (2014) 875–879.

[19] Z. Jiang, Z. Lin, and L. S. Davis, Label consistent k-svd: Learning a discriminative dictionary for recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013) 2651–2664.

[20] S.-J. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, An interior-point method for large-scale-regularized least squares, *IEEE J. Slected Topics Signal Processing* 1 (2007) 606–617.

[21] T.-K. Kim, J. Kittler and R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (2007) 1005–1018.

[22] N.X. Li, L. Pan, S. Zhou and H. Qi, Improved iterative hard thresholding for sparsity and nonnegativity constrained optimization, in http://www.personal.soton.ac.uk/hdqi/REPORTS/IIHT.pdf, 2016, pp.1–29.

[23] J. Liu, W. Liu, Q. Li, S. Ma and G. Chen, Evaluation of k-svd with different embedded sparse representation algorithms, in: *Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2016 12th International Conference on, 2016, pp. 426–432.

[24] Z. Lu, Optimization over sparse symmetric sets via a nonmonotone projected gradient method, arXiv preprint, 2015.

[25] A.M. Martinez, The ar face database, CVC Technical Report, 1998, vol.24.

[26] I. Naseem, R. Togneri and M. Bennamoun, Linear regression for face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (2010) 2106–2112.

[27] A. Rakotomamonjy, Surveyingandcomparingsimultaneoussparseapproximation(orgroup-lasso)algorithms, *Signal processing* 91 (2011) 1505–1526.

[28] N.M. Nasrabadi, T.D. Tran and N. Nguyen, Robust lasso with missing and grossly corrupted observations, in: *Advances in Neural Information Processing Systems*, 2011, pp. 1881–1889.

[29] A.V. Nefian, Georgia tech face database, 2013.

[30] D.-S. Pham and S. Venkatesh, Improved image recovery from compressed data contaminated with impulsive noise, *IEEE Transactions on Image Processing* 21 (2012) 397–405.

[31] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek, Overview of the face recognition grand challenge, in: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR' 05)*, vol. 1, 2005, pp. 947–954.

[32] Q. Shi, A. Eriksson, A. Van Den Hengel and C. Shen,in: *Is face recognition really a compressive sensing problem?*, *Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 553–560.

[33] Y. Su, S.Shan, X. Chen and W.Gao, Adaptive generic learning for face recognition from as ingle sample per person, in: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2699–2706.

[34] X. Tan, S. Chen, Z.-H. Zhou and F. Zhang, Face recognition from a single image per person: A survey, *Pattern recognition* 39 (2006) 1725–1745.

[35] J.A. Tropp, A.C. Gilbert and M.J. Strauss, Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit, *Signal Processing* 36 (2006) 572–588.

[36] R. Wang, H. Guo, L.S.Davis and Q.Dai, Covariance discriminative learning: A natural and efficient approach to image set classification, in: *Computer Vision and Pattern Recognition(CVPR)*, 2012, pp. 2496–2503.

[37] R. Wang, S. Shan, X. Chen and W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: *Computer Vision and Pattern Recognition(CVPR)*, 2008, pp. 1–8.

[38] J. Wright, A. Ganesh, A. Yang, Z. Zhou and Y. Ma, Sparsity and robustness in face recognition, 2011.

[39] J. Wright, A Y. Yang, A. Ganesh, S.S. Sastry and Y. Ma, Robust face recognition via sparse representation, *IEEE Tansactions on Pattern Analysis and Machine Intelligence* 31 (2009) 210–227.

[40] Z. Xu, X. Chang, F. Xu and H. Zhang, $L_{1/2}$ regularization: A thresholding representation theory and a fast solver, *IEEE Transactions on Neural Networks and Learning Systems* 23 (2012) 1013–1027.

[41] M. Yang, L. Van Gool and L. Zhang, Sparse variation dictionary learning for face recognition with a single training sample per person, in: *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 689–696.

[42] L. Yang, T.K. Pong and X. Chen, Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction, *SIAM J. Imaging Sci.* 10 (2017) 74–110.

[43] B. Zhang, S. Luan, C. Chen, J. Han, W. Wang, A. Perina and L. Shao, Latent constrained correlation filter, *IEEE Trans. Image Process* 27 (2018) 1038–1048.

[44] B. Zhang, A. Perina, C. Li, Q. Ye, V. Murino A. Del Bue, Manifold constraint transfer for visual structure-driven optimization, *Pattern Recognition* 77 (2018) 87–98.

[45] H. Zhang, N.M. Nasrabadi, Y. Zhang and T.S. Huang, Joint dynamic sparse representation for multi-view face recognition, *Pattern Recognition* 45 (2012) 1290–1298.

[46] Q. Zhang and B. Li, Discriminative k-svd fordictionary learning in face recognition, in: *Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2691–2698.

[47] X. Zhang, D.-S. Pham, S. Venkatesh, W. Liu and D. Phung, Mixed-norm sparse representation for multi view face recognition, *Pattern Recognition* 48 (2015) 2935–2946.

[48] W. Zhao, R.Chellappa, P.J. Phillips and A. Rosenfeld, Face recognition: A literature survey, *ACM computing surveys(CSUR)* 35 (2003) 399–458.

Manuscript received 11 April 2018
revised 26 October 2018
accepted for publication 20 November 2018

Jingjing Liu
State Key Laboratory of ASIC and System
Schoole of Microelectronics, Fudan University
Shanghai, P. R. China
School of Computer Science and Technology
Shanghai University of Electric Power
Shanghai, P. R. China
E-mail address: liujingjing@fudan.edu.cn

Shiwei Ma
School of Mechatronical Engineering and Automation
Shanghai University, Shanghai, P. R. China
E-mail address: masw@shu.edu.cn

Xianchao Xiu
Department of Mechanics and Engineering Science
Peking University, Beijing, P. R. China
E-mail address: xcxiu@bjtu.edu.cn