Calhoun: The NPS Institutional Archive

Faculty and Researcher Publications

Funded by Naval Postgraduate School

2016

Learning Robust and Discriminative Subspace With Low-Rank Constraints

Sheng Li

http://hdl.handle.net/10945/52406

Calhoun is a project of the Dudley Knox Library at NPS, furthering the precepts and goals of open government and government transparency. All information contained herein has been approved for release by the NPS Public Affairs Officer.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

http://www.nps.edu/library

Learning Robust and Discriminative Subspace With Low-Rank Constraints

Sheng Li, *Student Member, IEEE*, and Yun Fu, *Senior Member, IEEE*

***Abstract*— In this paper, we aim at learning robust and discriminative subspaces from noisy data. Subspace learning is widely used in extracting discriminative features for classification. However, when data are contaminated with severe noise, the performance of most existing subspace learning methods would be limited. Recent advances in low-rank modeling provide effective solutions for removing noise or outliers contained in sample sets, which motivates us to take advantage of low-rank constraints in order to exploit robust and discriminative subspace for classification. In particular, we present a discriminative subspace learning method called the supervised regularization-based robust subspace (SRRS) approach, by incorporating the low-rank constraint. SRRS seeks low-rank representations from the noisy data, and learns a discriminative subspace from the recovered clean data jointly. A supervised regularization function is designed to make use of the class label information, and therefore to enhance the discriminability of subspace. Our approach is formulated as a constrained rank-minimization problem. We design an inexact augmented Lagrange multiplier optimization algorithm to solve it. Unlike the existing sparse representation and low-rank learning methods, our approach learns a low-dimensional subspace from recovered data, and explicitly incorporates the supervised information. Our approach and some baselines are evaluated on the COIL-100, ALOI, Extended YaleB, FERET, AR, and KinFace databases. The experimental results demonstrate the effectiveness of our approach, especially when the data contain considerable noise or variations.**

***Index Terms*— Image classification, low-rank constraints, robust subspace discovery, subspace learning.**

I. Introduction

SUBSPACE learning methods have been extensively studied in pattern recognition and data mining areas during the last two decades. Some representative subspace learning methods include principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2], locality preserving projections (LPPs) [3], neighborhood preserving embedding (NPE) [4], locality sensitive discriminant analysis (LSDA) [5], and discriminative locality alignment (DLA) [6]. The basic idea of subspace learning methods is to find a low-dimensional projection that satisfies some specific properties [7]. As unsupervised methods, PCA [1] seeks such a subspace where the variance of projected samples is maximized, while LPP [3] and NPE [4] aim to find subspaces that can preserve the locality relationships of samples. When class labels are available, supervised subspace methods are more effective for classification tasks. LDA [2] aims at finding a projection that maximizes the interclass scatter and minimizes the intraclass scatter at the same time. It extracts discriminative features for classification tasks. LSDA [5] preserves both discriminant and local geometrical structure in data. DLA [6] is designed based on the patch alignment framework, which presents the idea of part optimization and whole alignment. As a discriminative model, it is suitable for the nonlinear classification problem. In [8], two generic frameworks are presented to implement supervised subspace learning for multilabel classification. Note that the frameworks built in [6] and [8] provide us with unified interpretations of many subspace learning methods. LPP [3] and NPE [4] can also be extended to supervised versions. Those methods usually obtain promising results on clean data; however, when the data are corrupted by considerable noise (e.g., missing pixels or outliers) or large variations (e.g., pose variations in face images) in real applications, their performance is heavily degraded [9].

To learn effective features from noisy data, many techniques have been introduced, and sparse representation (SR) is among the most successful ones. SR has proved to be robust to noise, and has shown impressive results for face recognition under noisy conditions [10], [11]. The idea of SR has also been considered in dimensionality reduction and subspace learning [12]–[15]. Zhang *et al.* [13] combine dimensionality reduction and an SR classifier. A sparsity preserving projections method is proposed in [12], and its improved version is introduced in [15]. Moreover, a linear subspace learning (LSL) algorithm via sparse coding is described in [14], which also involves dictionary learning. Most SR methods seek the sparsest coding vector to represent each test sample by all training samples. However, the underlying global structure of data is not considered in these methods, and therefore they may not be robust to noise when extra clean data are not available [16].

Low-rank modeling has attracted a lot of attention recently, which can recover the underlying structure of data [17], [18]. It is an extension of SR. When data are drawn from a

Manuscript received April 13, 2014; revised May 27, 2015 and July 1, 2015; accepted July 18, 2015. Date of publication August 31, 2015; date of current version October 17, 2016. This work was supported in part by the National Science Foundation Computer and Network Systems under Award 1314484, in part by the Office of Naval Research (ONR) under Award N00014-12-1-1028, in part by the ONR Young Investigator under Award N00014-14-1-0484, in part by the Naval Post-Graduate School under Award N00244-15-1-0041, and in part by the U.S. Army Research Office Young Investigator under Award W911NF-14-1-0218. *(Corresponding author: Yun Fu.)*

S. Li is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: lisheng1989@gmail.com).

Y. Fu is with the Department of Electrical and Computer Engineering, College of Computer and Information Science, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

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/TNNLS.2015.2464090

2162-237X © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

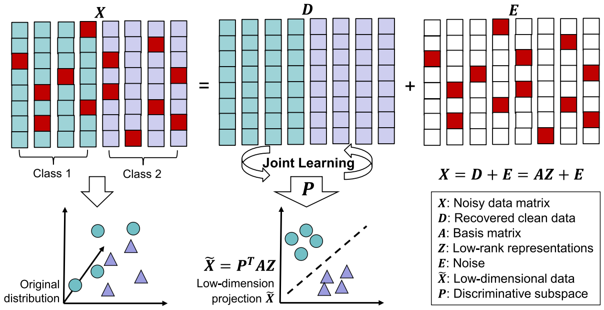

X
D
E
=
+
Class 1
Class 2
Joint Learning
P
$X = D + E = AZ + E$
Original distribution
$\widetilde{X} = P^T AZ$
Low-dimension projection $\widetilde{X}$
X: Noisy data matrix
D: Recovered clean data
A: Basis matrix
Z: Low-rank representations
E: Noise
$\widetilde{X}$: Low-dimensional data
P: Discriminative subspace

Fig. 1. Framework of the proposed approach. We jointly remove noise from data X and learn robust subspace P. The corrupted samples are mixed in the original space, but they are well separated in the learned subspace.

single subspace, robust PCA (RPCA) [17] is able to recover the corrupted data by minimizing the rank of data matrix. As an extension of RPCA, low-rank representation (LRR) [16] can recover corrupted data drawn from multiple subspaces. RPCA has been successfully applied to background modeling, and LRR achieves impressive performance on subspace clustering. Many improved versions of LRR have been developed. Latent LRR (LatLRR) [9] considers the effects of hidden data. Low-rank coding-based balanced graph is designed for clustering [19] and semisupervised classification [20]. In addition, low-rank modeling has been applied to outlier detection [21], domain adaptation [22], transfer learning [23], [24], and dictionary learning [25]–[27]. Low-rank modeling usually suffers large computational burden, and the idea of divide-and-conquer has been introduced to solve this problem [28], [29], which makes low-rank modeling scalable to larger data sets.

A. Our Contributions

As discussed above, the low-rank modeling has shown impressive performance in various applications [16], [17], [30]. However, only a few of those methods can take advantages of class label information during low-rank learning, which is a key for classification purpose. On the other hand, although the conventional subspace learning approaches usually obtain good performance for classification tasks, they have strong assumptions on the data distribution, and therefore, they are sensitive to the noisy data. The learned subspace has limited discriminability. Can we leverage the advantages of both supervised subspace learning and low-rank modeling for classification?

In this paper, we propose to exploit a discriminative and robust subspace, which is insensitive to noise or pose/illumination variations, for dimensionality reduction and classification. In particular, we propose a novel linear subspace approach named supervised regularization-based robust subspace (SRRS) for pattern classification. As shown in Fig. 1, the core idea of our approach is to jointly learn LRRs from the noisy data, and a discriminative subspace from the recovered clean data. Moreover, to improve the classification performance of our approach, we naturally incorporate class label information into our objective function as supervised regularization. This regularization term enables us to learn a discriminative subspace, which benefits classification tasks. Finally, we formulate our model as a constrained rank-minimization problem, and solve it using the recently proposed augmented Lagrange multiplier (ALM) algorithm [31]. The convexity of supervised regularization term is proved theoretically. The experimental results on six benchmark data sets show that our SRRS approach outperforms the traditional subspace methods and several state-of-the-art low-rank modeling methods in almost all cases, especially when the data contain considerable variations or are corrupted by noise.

This paper is a substantial extension of [32]. Compared with [32], we provide more theoretical analysis, model discussions, experimental evaluations, and applications in this paper. In summary, our contributions include the following.

1) We have proposed a new feature extraction framework, which smoothly integrates LSL and low-rank matrix recovery. Supervised regularization is incorporated to improve the classification performance.
2) We have designed an optimization algorithm to solve the proposed model, and have proven the convexity of the supervised regularization term.
3) Besides objection recognition and face recognition evaluated in [32], we have also extended the applications of our model to kinship verification. We have also provided a more comprehensive overview of related works.

The rest of this paper is organized as follows. We briefly introduce some related works in Section II. Then, we describe the proposed SRRS approach, theoretical analysis, and optimization algorithm in Section III. Experiments are reported in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Work

In this section, we will review two categories of related methods: 1) subspace learning and 2) low-rank modeling.

A. Subspace Learning

Subspace learning has been extensively studied and widely used in many real-world applications, such as face recognition, object recognition, and visualization. The basic idea of subspace learning methods is to project high-dimensional samples into a low-dimensional subspace, in which some specific properties could be satisfied. According to the usage of class labels, subspace learning methods are mainly divided into three categories: 1) unsupervised methods; 2) supervised methods; and 3) semisupervised methods. In this paper, we only focus on the supervised ones.

Supervised subspace learning methods are very effective in extracting discriminative features, and usually achieve promising performance in classification tasks. LDA [2] is developed upon the Fisher criterion, which aims at finding a projection to maximize the interclass scatter and minimize the intraclass scatter simultaneously. Many supervised subspace methods have been proposed to improve LDA. Local Fisher discriminant analysis (LFDA) [33] uses local neighborhood information to construct the weighted between-class and within-class scatter matrices, and then performs discriminant analysis. Subclass discriminant analysis [34] models the data using mixture of Gaussians, and redefines the scatter matrices used in LDA. LSDA [5] preserves both discriminant and local geometrical structure in data. Those methods usually obtain promising results on clean data, since they place specific assumptions on data distributions. However, when the data are corrupted by large amount of noise or large variations in real applications, these assumptions may be invalid, and the noise or variation can reduce the separability in a classification task. Therefore, the performance is heavily degraded.

B. Low-Rank Modeling

Low-rank modeling is becoming popular and practical recently [35], due to its successful applications in many fields, such as data compression [17], subspace clustering [16], [36], image processing [37], [38], and multimedia analysis [39]. RPCA [17] is a representative low-rank modeling method. Given an observed and usually corrupted sample set X_O, RPCA decomposes X_O into a low rank, clean sample set X_L and a sparse, noisy sample set E, i.e., $X_O = X_L + E$. It shows impressive performance in background modeling and shadow removal. One major assumption in RPCA is that data are drawn from a single subspace. In practice, the underlying structure of data could be multiple subspaces. LRR is designed to find underlying structures of noisy data [16]. Given a sample set $X = [x_1, x_2, \ldots, x_n]$, the objective function of LRR is

$$\min_{Z} \ \mathrm{rank}(Z), \quad \text{s.t. } X = XZ \tag{1}$$

where Z is the representation coefficient matrix, and the sample set X is used as the dictionary. By replacing the rank$(\cdot)$ function with nuclear norm, problem (1) can be converted into a convex optimization problem.

The LRR may suffer from two problems. The first one is insufficient data sampling since LRR simply uses the data matrix itself as the basis for representation. Second, the optimization of LRR requires multiple singular value decomposition (SVD) calculations that are very time consuming. In [9], LatLRR is proposed to solve the insufficient sampling problem by considering the effects of hidden data for representation. In addition, active subspace [40] and divide-factor-combine LRR [29] employ various matrix factorization algorithms to tackle the above problems. The aim of these low-rank modeling methods is to learn a graph Z, and they do not utilize any supervised information. However, our goal is to learn a discriminative low-dimensional subspace. Although subspace learning methods can be combined with LatLRR [9], the representation learnt by the LatLRR does not necessarily guarantee an optimal input for the subsequent subspace learning. Nevertheless, our approach simultaneously seeks optimal LRRs and discriminative subspaces.

The discriminative low-rank dictionary learning (DLRD) [25] is a recently proposed dictionary learning method, which introduces low-rank constraints on the subdictionaries for each class, and performs SR for face recognition. The learned dictionary in DLRD is low-rank and discriminative, which is beneficial to classification tasks. Nevertheless, the testing stage of DLRD is very time consuming, as it has to calculate sparse coefficients for every test sample. This is also a key difference between DLRD and our approach, since we perform classification on subspace that is very efficient.

In [41], a low-rank method with structural incoherence is applied to face recognition. It first decomposes raw images into low-rank part and sparse part, and then applies PCA on the low-rank part to obtain a subspace. Finally, it employs SR for classification. They did not, however, learn the LRR and a discriminative subspace simultaneously. In this manner, the low-rank part is expected to be discriminative and benefit classification tasks.

In [26], a structured LRR method is presented for image classification. The differences between [26] and our approach include the following.

1) Zhang *et al.* [26] learned a dictionary D to represent the sample set X in the original sample space, but our approach aims at learning a low-dimensional discriminative subspace to reduce the dimensionality of samples.
2) Zhang *et al.* [26] enforced a diagonal structure prior on the coefficient matrix Z to introduce the supervised information, but our approach employs the Fisher criterion to learn discriminative features.
3) Zhang *et al.* [26] used the ridge regression model for classifying new samples, but our approach adopts the nearest neighbor (NN) classifier.

In [42], an LRR-based discriminative projection method is proposed for feature extraction. It first applies LRR to recover the data matrix, and then finds a discriminative projection by designing a criterion that incorporates both clean data and noise. In this case, LRR is regarded as a data preprocessing method, and is performed only once to decompose sample set into two parts: 1) the low-rank denoised samples and 2) the associated sparse noise. However, this decomposition is not guaranteed to be optimal for classification, as it does not make use of any class prior information. On the contrary, our approach iteratively learns subspace and decomposes sample set, and it takes full advantage of class information through supervised regularization.

In [43], a discriminant regularization term is incorporated into the formulation of RPCA. This method differs from our approach in two aspects. First, RPCA used in [43] can only model one single subspace, whereas our approach is able to discover multiple subspaces by virtue of LRR, which fits well for multiclass classification problems. Second, the method in [43] separately learns low-rank data representation and subspace, which means the obtained subspace cannot be guaranteed to be optimal, whereas our approach iteratively learns LRRs and discriminative subspaces.

The most relevant method in the literature is low-rank transfer subspace learning (LTSL) [23], [24], which incorporates low-rank constraint in subspace learning. However, there are significant differences between LTSL and our approach. First, the LTSL is a transfer learning method that seeks a common subspace for two domains, whereas our approach lies in supervised learning. Second, the LTSL employs low-rank constraint in low-dimensional subspace in order to transfer knowledge across two domains. In our approach, the low-rank constraint is enforced in the high-dimensional feature space in order to preserve more information.

III. Supervised Regularization-Based Robust Subspace Approach

In this section, an SRRS approach is proposed. We first formulate our approach as a regularized rank-minimization problem. To solve this problem, we develop an efficient optimization algorithm. The theoretical analysis on convexity is also provided.

A. Problem Formulation

Let X denote the sample set that consists of n training samples from c classes, i.e., $X = [x_1, x_2, \ldots, x_n]$. Given a complete basis matrix $A = [a_1, a_2, \ldots, a_m] \in \mathbb{R}^{d\times m}$, we can represent each sample x_i as a linear combination of the basis, which is

$$X = AZ \tag{2}$$

where $Z \in \mathbb{R}^{m\times n}$ is the coefficient matrix. As suggested in the existing subspace clustering methods, A is usually set as the sample set X, i.e., $A = X$. We will discuss the choice of basis matrix A at the end of this section.

To achieve our goal of seeking a robust subspace $P \in \mathbb{R}^{d\times p}$, we first denote the projected low-dimensional sample set as $\tilde{X} = P^{\mathrm{T}}X = P^{\mathrm{T}}AZ$. Then, we in turn incorporate low-rank constraint and supervised regularization to learn the projection P.

First, due to the fact that n samples belong to c different classes and $n \gg c$, these samples should be drawn from c different subspaces, and therefore, the coefficient matrix Z is expected to be low rank. In other words, the coefficient vectors corresponding to samples from the same class should be highly correlated.

Second, since class information is crucial to classification problems, we design a supervised regularization term $f(P, Z)$ based on the idea of Fisher criterion [2], that is, $f(P, Z) = [\mathrm{Tr}(S_B(P^{\mathrm{T}}AZ))/\mathrm{Tr}(S_W(P^{\mathrm{T}}AZ))]$, where $\mathrm{Tr}(K)$ is the trace of matrix K. $S_B(P^{\mathrm{T}}AZ)$ and $S_W(P^{\mathrm{T}}AZ)$ are the between-class and within-class scatter matrices

$$S_B(P^{\mathrm{T}}AZ) = S_B(\tilde{X}) = \sum_{i=1}^{c} n_i(m_i - m)(m_i - m)^{\mathrm{T}}$$
$$S_W(P^{\mathrm{T}}AZ) = S_W(\tilde{X}) = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(\tilde{x}_{ij} - m_i)(\tilde{x}_{ij} - m_i)^{\mathrm{T}}$$

where m_i is the mean sample of the ith class in $\tilde{X}$, m is the overall mean sample of $\tilde{X}$, and $\tilde{x}_{ij}$ is the jth sample in the ith class of $\tilde{X}$.

By using Fisher criterion, the projected samples from different classes should be far apart, while projected samples from the same class should be close to each other. Furthermore, Guo *et al.* [44] pointed out that this trace-ratio problem can be converted into a trace difference problem. We then rewrite $f(P, Z)$ as $\bar{f}(P, Z) = \mathrm{Tr}(S_W(P^{\mathrm{T}}AZ)) - \mathrm{Tr}(S_B(P^{\mathrm{T}}AZ))$.

Based on the above observations, we come up with the following objective function:

$$\min_{Z,P}\ \mathrm{rank}(Z) + \lambda_1\bar{f}(P, Z), \quad \text{s.t. } X = AZ \tag{3}$$

where λ_1 is a tradeoff parameter to balance the low rank and discriminative terms.

However, the rank-minimization problem in objective (3) is difficult to solve, since $\mathrm{rank}(\cdot)$ is a nonconvex function. Fortunately, nuclear norm is a good surrogate for the rank-minimization problem [16], [17], [45], and then (3) becomes

$$\min_{Z,P}\ \|Z\|_* + \lambda_1\bar{f}(P, Z), \quad \text{s.t. } X = AZ \tag{4}$$

where $\|Z\|_*$ is the nuclear norm of a matrix (i.e., the sum of singular values of the matrix) [46].

We also notice that the second term $\bar{f}(P, Z)$ in (4) is not convex to Z because of the term $-\mathrm{Tr}(S_B)$, so we add an elastic term to ensure the convexity

$$\hat{f}(P, Z) = \mathrm{Tr}(S_W) - \mathrm{Tr}(S_B) + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2. \tag{5}$$

We theoretically prove the convexity of (5) in Section III-B.

Equation (5) can be equivalently expressed as

$$\hat{f}(P, Z) = \|P^{\mathrm{T}}AZ(\mathbf{I} - H_b)\|_{\mathrm{F}}^2 - \|P^{\mathrm{T}}AZ(H_b - H_t)\|_{\mathrm{F}}^2 + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2 \tag{6}$$

where η is a tradeoff parameter, $\|.\|_{\mathrm{F}}$ is the Frobenius norm, $\mathbf{I}$ is an identity matrix in $\mathbb{R}^{n\times n}$, and H_b and H_t are two constant coefficient matrices. In detail, $H_b(i, j) = (1/n_c)$ only if x_i and x_j belong to the same class, where n_c is the number of samples in each class; otherwise, $H_b(i, j) = 0$. $H_t(i, j) = (1/n)$.

The supervised regularization term $\hat{f}(P, Z)$ is convex with respect to Z. We will provide the theoretical analysis to prove it in Section III-B.

Orthogonality in a subspace means that any two basis vectors in this subspace are orthogonal to each other, which has the advantages of compactness and reducing redundancy. To this end, an orthogonal constraint $P^{\mathrm{T}}P = \mathbf{I}_p$ is incorporated into our framework, where $\mathbf{I}_p$ is an identity matrix in $\mathbb{R}^{p\times p}$. By combining (4) and (6), we obtain the objective function as follows:

$$\begin{aligned}\min_{Z,P}\ & \|Z\|_* + \lambda_1\big(\|P^{\mathrm{T}}AZ(\mathbf{I} - H_b)\|_{\mathrm{F}}^2 \\ & - \|P^{\mathrm{T}}AZ(H_b - H_t)\|_{\mathrm{F}}^2 + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2\big) \\ \text{s.t. } & X = AZ, \quad P^{\mathrm{T}}P = \mathbf{I}_p.\end{aligned} \tag{7}$$

Note that our objective function in (7) is not convex with respect to P, because of the orthogonal constraint $P^{\mathrm{T}}P = \mathbf{I}_p$.

In real-world applications, as we discussed in Section I, data usually contain considerable noise. To obtain robust subspaces, we should identify noisy information in raw data, and learn reliable subspaces from the recovered noise-free data. In particular, we adopt the $l_{2,1}$-norm (i.e., $\|\cdot\|_{2,1}$) to model the noise contained in data. $l_{2,1}$-norm is a valid norm as it satisfies three conditions for a norm: 1) positive scalability: $\|\alpha E\|_{2,1} = |\alpha|\|E\|_{2,1}$, where α is a real scalar; 2) triangle inequality: $\|B+E\|_{2,1} \leq \|B\|_{2,1} + \|E\|_{2,1}$; and 3) existence of a zero vector: if $\|E\|_{2,1} = 0$, then $A = 0$. As $\|E\|_{2,1}$ encourages the columns of E to be zero, the assumption in this paper is that some vectors in our data are corrupted, while the others are clean. Then, we have a constraint $X = AZ + E$, and rewrite the objective function as

$$\begin{aligned}\min_{Z,E,P}\ & \|Z\|_* + \lambda_2\|E\|_{2,1} \\ & + \lambda_1\big(\|P^{\mathrm{T}}AZ(\mathbf{I}-H_b)\|_{\mathrm{F}}^2 \\ & \quad - \|P^{\mathrm{T}}AZ(H_b-H_t)\|_{\mathrm{F}}^2 + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2\big) \\ \text{s.t. } & X = AZ+E, \quad P^{\mathrm{T}}P = \mathbf{I}_p \end{aligned} \tag{8}$$

where $\|E\|_{2,1} = \sum_{j=1}^{n}\big(\sum_{i=1}^{d}([E]_{ij})^2\big)^{1/2}$, and λ_2 is a tradeoff parameter.

We have described how to jointly learn discriminative subspace and LRRs. In Section III-B, we will introduce the optimization algorithm. Other than Fisher criterion discussed above, other types of objectives, such as locality preserving, can also be easily incorporated into our framework by reformulating the regularization term $\hat{f}(P, Z)$.

B. Theoretical Analysis

We theoretically analyze the convexity of supervised regularization term $\hat{f}(P, Z)$ with respect to Z, which is critical to ensure that our model is solvable using ALM algorithms. In particular, to guarantee the convexity of (6), we provide Theorem 1.

Theorem 1: If $\eta > 1$, the supervised regularization term $\hat{f}(P, Z) = \|P^{\mathrm{T}}AZ(\mathbf{I}-H_b)\|_{\mathrm{F}}^2 - \|P^{\mathrm{T}}AZ(H_b-H_t)\|_{\mathrm{F}}^2 + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2$ is convex to Z when P is fixed.

Proof: Let $T = P^{\mathrm{T}}AZ$, where $P^{\mathrm{T}}A$ can be regarded as constant when optimizing Z. We then can convert $\hat{f}(P, Z)$ to $f(T)$ as follows:

$$f(T) = \|T(\mathbf{I}-H_b)\|_{\mathrm{F}}^2 - \|T(H_b-H_t)\|_{\mathrm{F}}^2 + \eta\|T\|_{\mathrm{F}}^2. \tag{9}$$

Now, we can rewrite T as a column vector, $\mathbf{T} = [r_1, r_2, \ldots, r_n]^{\mathrm{T}}$, where r_i is the ith row vector of T. Then, $f(T)$ is equivalent to

$$\begin{aligned} f(\mathbf{T}) = & \|\mathrm{diag}((\mathbf{I}-H_b)^{\mathrm{T}})\mathbf{T}\|_2^2 - \|\mathrm{diag}((H_b-H_t)^{\mathrm{T}})\mathbf{T}\|_{\mathbf{2}}^{\mathbf{2}} \\ & + \eta\|\mathbf{T}\|_2^2 \end{aligned} \tag{10}$$

where $\mathrm{diag}(K)$ is to construct a block diagonal matrix with each block on the diagonal being matrix K.

The convexity of $f(\mathbf{T})$ depends on whether its Hessian matrix $\nabla^2 f(\mathbf{T})$ is positive definite or not. $\nabla^2 f(\mathbf{T})$ will be positive definite if matrix S is positive definite

$$S = (\mathbf{I}-H_b)(\mathbf{I}-H_b)^{\mathrm{T}} - (H_b-H_t)(H_b-H_t)^{\mathrm{T}} + \eta\mathbf{I}. \tag{11}$$

Note that we have the equations $H_bH_t = H_tH_b = H_t$ and $H_tH_t = H_t$. Then, we can obtain

$$S = (1+\eta)\mathbf{I} - 2H_b + H_t. \tag{12}$$

To justify that if matrix S is positive definite, we employ Lemma 1.

Lemma 1 (Weyl's Inequality [47, Th. 1]): Let G denote an $n \times n$ Hermitian matrix, the ordered eigenvalues of G are $\lambda_1(G) \geq \cdots \geq \lambda_n(G)$. If B and C are $n \times n$ Hermitian matrices, then $\lambda_n(B) + \lambda_n(C) \leq \lambda_n(B+C)$.

Lemma 1 tells us the smallest eigenvalue of matrix $(B+C)$ is greater than or equal to the sum of the smallest eigenvalues of B and C. In our problem, we need to make S positive definite, which means the smallest eigenvalue of S should be greater than 0. Thus, we employ Lemma 1 to evaluate the (12). The minimal eigenvalues of $-H_b$ and H_t are -1 and 0, so we should ensure

$$(1+\eta) - 2 + 0 > 0. \tag{13}$$

Hence, we have $\eta > 1$ from (13), which could guarantee that $f(T)$ is convex to T. Recall that $T = P^{\mathrm{T}}AZ$ and $P^{\mathrm{T}}A$ is a constant. Therefore, we can further conclude that $f(P, Z)$ is convex to Z when $\eta > 1$ and P is fixed. □

C. Optimization

To solve (8), we adopt the recently proposed inexact ALM algorithm [31]. First, we add a variable J and a new constraint $Z = J$ to relax the original problem

$$\begin{aligned}\min_{Z,E,P,J}\ & \|J\|_* + \lambda_2\|E\|_{2,1} \\ & + \lambda_1\big(\|P^{\mathrm{T}}AZ(\mathbf{I}-H_b)\|_{\mathrm{F}}^2 \\ & \quad - \|P^{\mathrm{T}}AZ(H_b-H_t)\|_{\mathrm{F}}^2 + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2\big) \\ \text{s.t. } & X = AZ+E, \quad P^{\mathrm{T}}P = \mathbf{I}_p, \quad Z = J. \end{aligned} \tag{14}$$

Furthermore, (14) can be converted into the following problem:

$$\begin{aligned}\min_{Z,E,J,P,Y,R}\ & \|J\|_* + \lambda_2\|E\|_{2,1} \\ & + \lambda_1\big(\|P^{\mathrm{T}}AZ(\mathbf{I}-H_b)\|_{\mathrm{F}}^2 \\ & \quad - \|P^{\mathrm{T}}AZ(H_b-H_t)\|_{\mathrm{F}}^2 + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2\big) \\ & + \mathrm{Tr}(Y^T(X-AZ-E)) + \mathrm{Tr}(R^T(Z-J)) \\ & + \frac{\mu}{2}\big(\|X-AZ-E\|_{\mathrm{F}}^2 + \|Z-J\|_{\mathrm{F}}^2\big) \\ \text{s.t. } & P^{\mathrm{T}}P = \mathbf{I}_p \end{aligned} \tag{15}$$

where $\mu > 0$ is a penalty parameter and $Y \in \Re^{d\times n}$ and $R \in \Re^{m\times n}$ are Lagrange multipliers.

To solve (15), we alternately update the variables P, J, Z, and E. First, we learn a subspace P given an initialized LRR matrix Z. Second, on the fixed subspace P, we update the LRR matrix J and Z and the noise matrix E. Although the convergence of inexact ALM algorithm cannot be guaranteed when there are three or more variables, some theoretical results have been presented to ensure the convergence with mild conditions [16]. In addition, we demonstrate the convergence properties of our algorithm in the experiments.

1) Learn Subspace P on Fixed Low-Rank Representations: We first discuss how to optimize P while fixing Z, J, and E. Note that $\|J\|_* + \lambda_2\|E\|_{2,1} + \mathrm{Tr}(Y^{\mathrm{T}}(X-AZ-E)) + \mathrm{Tr}(R^{\mathrm{T}}(Z-J)) + (\mu/2)(\|X-AZ-E\|_{\mathrm{F}}^2 + \|Z-J\|_{\mathrm{F}}^2)$ can be regarded as constant.

The objective function with respect to P becomes

$$\begin{aligned} P_{k+1} = \min_{P_k} \ & \lambda_1\big(\big\|P_k^{\mathrm{T}}AZ_k(\mathbf{I}-H_b)\big\|_{\mathrm{F}}^2 \\ & - \big\|P_k^{\mathrm{T}}AZ_k(H_b-H_t)\big\|_{\mathrm{F}}^2 + \eta\big\|P_k^{\mathrm{T}}AZ_k\big\|_{\mathrm{F}}^2\big) \\ \text{s.t. } & P_k^{\mathrm{T}}P_k = \mathbf{I}_p. \end{aligned} \tag{16}$$

For simplicity, let $Z_{wk} = AZ_k(\mathbf{I}-H_b)$ and $Z_{bk} = AZ_k(H_b-H_t)$. We derive the solution to the projection vectors in P_k one-by-one. To obtain the ith column in P_k [denoted as $P_{k(:,i)}$], we rewrite (16) as

$$\begin{aligned} P_{k+1(:,i)} = \min_{P_{k(:,i)}} \ & \lambda_1\big(\big\|P_{k(:,i)}^{\mathrm{T}}Z_{wk}\big\|_2^2 - \big\|P_{k(:,i)}^{\mathrm{T}}Z_{bk}\big\|_2^2 \\ & + \eta\big\|P_{k(:,i)}^{\mathrm{T}}AZ_k\big\|_2^2\big) \\ & + \beta_i\big(P_{k(:,i)}^{\mathrm{T}}P_{k(:,i)} - 1\big) \end{aligned} \tag{17}$$

where β_i is the corresponding Lagrange multiplier.

By setting the derivative with respect to $P_{k(:,i)}$ to zero, we have

$$-\lambda_1\big(Z_{wk}Z_{wk}^{\mathrm{T}} - Z_{bk}Z_{bk}^{\mathrm{T}} + \eta AZ_kZ_k^{\mathrm{T}}A^{\mathrm{T}}\big)P_{k(:,i)} = \beta_i P_{k(:,i)}. \tag{18}$$

Therefore, $P_{k(:,i)}$ is the ith eigenvector of matrix $-\lambda_1(Z_{wk}Z_{wk}^{\mathrm{T}} - Z_{bk}Z_{bk}^{\mathrm{T}} + \eta AZ_kZ_k^{\mathrm{T}}A^{\mathrm{T}})$, corresponding to the ith smallest eigenvalue.

2) Learn Low-Rank Representations Z on Fixed Subspace: Here, we show how to update J_{k+1}, Z_{k+1}, and E_{k+1} when fixing P_{k+1}. After dropping the irrelevant terms with respect to J, (15) can be rewritten as

$$\begin{aligned} J_{k+1} &= \min_{J_k} \ \|J_k\|_* + \mathrm{Tr}(R^{\mathrm{T}}(Z_k - J_k)) + \frac{\mu_k}{2}\|Z_k - J_k\|_{\mathrm{F}}^2 \\ &= \min_{J_k} \ \frac{1}{\mu_k}\|J_k\|_* + \frac{1}{2}\|J_k - (Z_k + (R_k/\mu_k))\|_{\mathrm{F}}^2. \end{aligned} \tag{19}$$

Problem (19) can be effectively solved using the singular value thresholding (SVT) operator [46]. SVT contains two major steps. First, we perform SVD on the matrix S ($S = Z_k + (R_k/\mu_k)$), and get $S = U_S\Sigma_SV_S$, where $\Sigma_S = \mathrm{diag}(\{\sigma_i\}_{1\le i\le r})$, σ_i is the singular value with rank r. Second, we can obtain the optimal solution J_{k+1} by thresholding the singular values: $J_{k+1} = U_S\Omega_{(1/\mu_k)}(\Sigma_S)V_S$, where $\Omega_{(1/\mu_k)}(\Sigma_S) = \mathrm{diag}(\{\sigma_i - (1/\mu_k)\}_+)$, and t_+ means the positive part of t.

By ignoring terms independent of Z in (15), we have

$$\begin{aligned} \min_{Z,Y,R} \ & \lambda_1\big(\|P^{\mathrm{T}}AZ(\mathbf{I}-H_b)\|_{\mathrm{F}}^2 - \|P^{\mathrm{T}}AZ(H_b-H_t)\|_{\mathrm{F}}^2 \\ & + \eta\|P^{\mathrm{T}}AZ\|_{\mathrm{F}}^2\big) + \mathrm{Tr}(Y^{\mathrm{T}}(X - AZ - E)) \\ & + \mathrm{Tr}(R^{\mathrm{T}}(Z-J)) \\ & + \frac{\mu}{2}\big(\|X - AZ - E\|_{\mathrm{F}}^2 + \|Z - J\|_{\mathrm{F}}^2\big). \end{aligned} \tag{20}$$

By setting the derivative with respect to Z to zero, we have

$$\begin{aligned} Z_{k+1}D/\mu_k + \big(A^{\mathrm{T}}P_{k+1}P_{k+1}^{\mathrm{T}}A\big)^{-1}(\mathbf{I} + A^{\mathrm{T}}A)Z_{k+1} \\ = \big(A^{\mathrm{T}}P_{k+1}P_{k+1}^{\mathrm{T}}A\big)^{-1}K_{k+1} \end{aligned} \tag{21}$$

where $D = \lambda_1((1+\eta)\mathbf{I} + H_bH_t{}^{\mathrm{T}} - H_b - H_tH_t{}^{\mathrm{T}})$, and $K_{k+1} = J_{k+1} + A^{\mathrm{T}}(X - E_k) + (A^{\mathrm{T}}Y_k - R_k)/\mu_k$. Problem (21) is a standard Sylvester equation, which can be effectively solved using the existing tools [48].

Algorithm 1 Solving Problem (15) by Inexact ALM

```
Input: data matrix $X$, parameter $\lambda_1$, $\lambda_2$, $\eta$, $Z = J = 0$,
    $E_0 = 0$, $Y_0 = 0$, $R_0 = 0$, $\mu_0 = 0.1$, $\mu_{\max} = 10^{10}$,
    $\rho = 1.3$, $k = 0$, $\epsilon = 10^{-8}$
Output: $P_k$, $Z_k$, $E_k$
1: while not converged do
2:   update $P_{k+1}$ using (17), given others fixed
     If $k = 1$, then $Z_k = \mathbf{I}$.
3:   update $J_{k+1}$ using (19), given others fixed
4:   update $Z_{k+1}$ using (21), given others fixed
5:   update $E_{k+1}$ using (23), given others fixed
6:   update the multipliers $Y_{k+1}$ and $R_{k+1}$
     $Y_{k+1} = Y_k + \mu_k(X - AZ_{k+1} - E_{k+1})$
     $R_{k+1} = R_k + \mu_k(Z_{k+1} - J_{k+1})$
7:   update the parameter $\mu_{k+1}$ by
     $\mu_{k+1} = \min(\rho\mu_k, \mu_{\max})$
8:   check the convergence conditions
     $\|X - AZ_{k+1} - E_{k+1}\|_\infty < \epsilon$   and
     $\|Z_{k+1} - J_{k+1}\|_\infty < \epsilon$.
9:   $k = k + 1$
10: end while
```

Similarly, after dropping terms independent of E, we can rewrite (15) as

$$\begin{aligned} E_{k+1} = \min_{E_k} \ & \frac{\lambda_2}{\mu_k}\|E_k\|_{2,1} \\ & + \frac{1}{2}\|E_k - (X - AZ_{k+1} + Y_k/\mu_k)\|_{\mathrm{F}}^2. \end{aligned} \tag{22}$$

The solution to problem (22) is presented in [16]. In particular, let $\Psi = X - AZ_{k+1} + Y_k/\mu_k$, the ith column of E_{k+1} is

$$E_{k+1}(:,i) = \begin{cases} \dfrac{\|\Psi_i\| - \lambda_2}{\|\Psi_i\|}\Psi_i, & \text{if } \lambda_2 < \|\Psi_i\| \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

As stated in the inexact ALM algorithm, we also need to update the Lagrange multipliers Y and R, and the parameter μ after optimizing the variables P, J, Z, and E.

D. Algorithm and Discussion

The above process is repeated until convergence. The detailed algorithm of our optimization is outlined in Algorithm 1.

After obtaining the optimal solution P^* and Z^*, we project both training samples and test samples onto P^*, and then utilize NN classifier to predict the label vector of test samples. The complete procedures of our SRRS approach are summarized in Algorithm 2.

The time complexity of our approach mainly depends on the complexity of Algorithm 1. In Algorithm 1, the most time-consuming steps are Steps 2–4. Steps 2 and 4 cost $O(n^3)$ due to the SVD decomposition, where n is the total number of samples. The matrix inverse calculation in (21) costs $O(n^3)$, and the state-of-the-art solution to a Sylvester equation costs $O(n^3 + m^3)$ (in our case, $m = n$). In all, the overall time

Algorithm 2 SRRS Approach

```
Input: Training sample set $X$ with label vectors $L_X$,
    test sample set $Y$, low-rank coefficients $Z$
Output: Predicted label vector $L_Y$ for test samples.
  1: Normalize each sample $x_i$ to unit-norm,
     $x_i = x_i / \|x_i\|$.
  2: Use Algorithm 1 to solve problem (15) and
     obtain optimal solution $P^*$.
  3: Project $X$ and $Y$ onto $P^*$:
     $\tilde{X} = P^{*\mathrm{T}} XZ$, $\tilde{Y} = P^{*\mathrm{T}} Y$.
  4: Predict the label vector $L_Y$ of $\bar{Y}$ by using the
     nearest neighbor (NN) classifier
```

complexity of our approach is $O(tn^3)$, where t is the number of iterations.

Formula (4) is actually a general framework for robust subspace learning and feature extraction. In this paper, we design a supervised regularization term $\bar{f}(P, Z)$ by virtue of Fisher criterion. Other subspace learning baselines (e.g., LPP, NPE, and LFDA) could also be extended under our framework by reformulating the regularization term $\bar{f}(P, Z)$.

In Algorithms 1 and 2, sample set X is utilized as dictionary (i.e., $A = X$). When the sampling is insufficient, learning an informative dictionary should enhance the classification performance, which provides another interesting direction of the future work.

In Step 3 of Algorithm 2, we project the recovered clean training images XZ onto the subspace P. Ideally, we would also like to project the clean test images onto P for classification. However, it is usually not practical to obtain clean test images in real applications. In this paper, to show the robustness of P for noisy data, we directly project noisy images onto P. To enhance the classification performance, one could apply some image denoising techniques before projecting noisy test data onto P.

IV. Experiments

The performance of our SRRS approach is evaluated on six benchmark data sets, including object data sets [49], [50], face data sets [51], [52], and KinFace data set [53]. We compare our approach with related methods on the robustness to different types of noise, including pixel corruption and large pose/illumination variations. Our code is publicly available.[1]

A. Object Recognition With Pixel Corruption

We use two object data sets, COIL-100 [49] and ALOI [50], in this experiment. The COIL data set contains various views of 100 objects with different lighting conditions. Each object contributes 72 images, which are captured in equally spaced views. In our experiments, the images are converted to grayscale, resized to 32 × 32, and then, the robustness is evaluated on alternative viewpoints. We normalize the samples so that they have unit norm that is favorable for optimization. Unlike the most existing subspace learning experiments, we also test the robustness of different methods to noise by adding 10% pixel corruption to the original images. Some examples of corrupted object images in COIL data set can be found in Fig. 9.

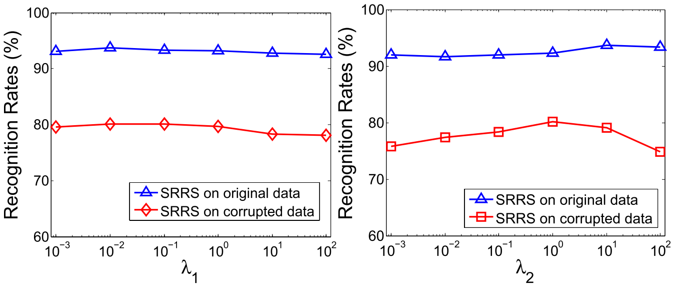


Fig. 2. Recognition rates of SRRS with different values of λ_1 and λ_2 on COIL data set.

In the experiments, we compare the proposed approach with PCA [1], LDA [2], NPE [4], LSDA [5], RPCA [17] + LDA, support vector machine (SVM) [54], FDDL [55], LatLRR [9], and DLRD [25]. PCA and LDA are two representative unsupervised and supervised subspace learning methods, and we use them as our baseline. NPE can preserve the neighborhood structure of data, which is less sensitive to outliers than PCA. Here, we compare our method with the supervised version of NPE. LSDA is a discriminant analysis method that preserves both discriminant and local geometrical structural in the data. RPCA is effective in removing noise from corrupted data. Here, we incorporate it with LDA as a baseline. SVM is a popular and powerful classifier. Here, we compared with the nonlinear SVM classifier with RBF kernel. FDDL is a dictionary learning method that learns a discriminative dictionary using Fisher criterion. LatLRR and DLRD are two low-rank modeling methods. LatLRR can effectively extract salient features for image recognition, whereas DLRD learns a low-rank dictionary for face recognition. Both of them also demonstrate stable performance under noisy conditions.

We randomly select ten images per object to construct the training set, and the test set contains the rest of the images. This random selection process is repeated 20 times, and we report the average recognition rates for each compared method. In addition, we performed scalability evaluations, by increasing the number of objects from 20 to 100. For our approach and each compared method, the parameters are tuned to achieve their best performance via fivefold cross validation. Fig. 2 shows the performance of SRRS with different values of λ_1 and λ_2 when the number of classes is 20. We can also observe that the performance is not very sensitive to the settings of λ_2. Furthermore, SRRS obtains its best performance when $\lambda_1 = 10^{-1}$. It also shows that the SRRS achieves the best performance on the original data and corrupted data when $\lambda_2 = 10$ and $\lambda_2 = 1$, respectively.

Fig. 3 shows the recognition rates of our approach and the compared subspace methods (PCA, LDA, NPE, and LSDA) versus varying feature dimensions. It shows that our SRRS approach outperforms subspace methods in almost all cases. When the images contain noise, the recognition rates of compared subspace methods are severely degraded, but our approach can still obtain good results. Namely, the subspace

[1] https://github.com/smilesheng/SRRS

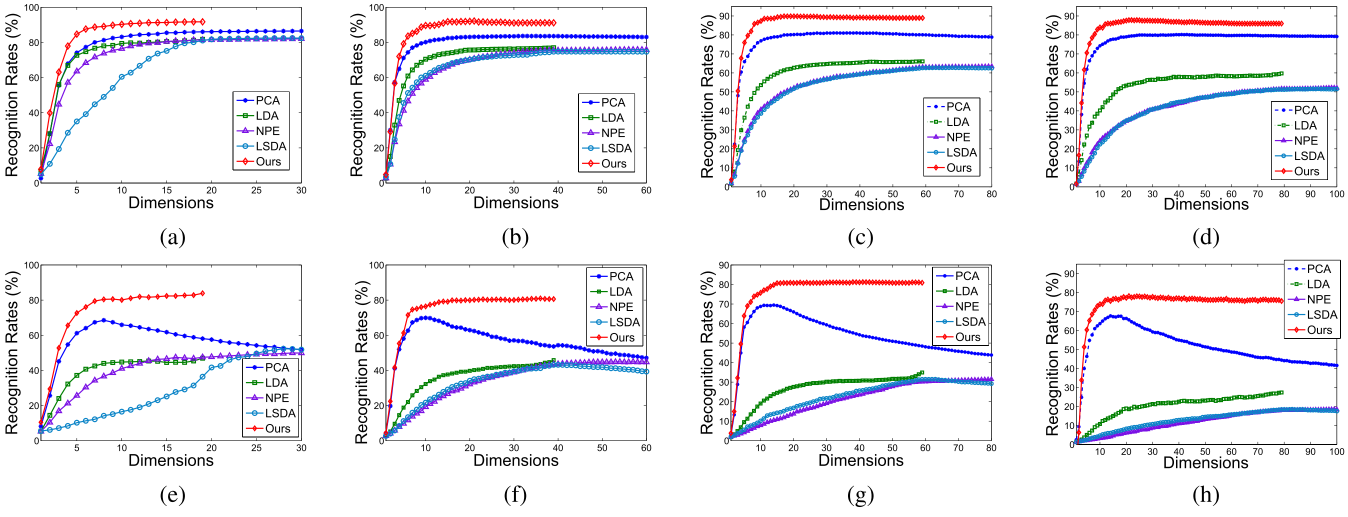


Fig. 3. Recognition rates of our approach and compared subspace methods versus varying feature dimensions on the original and corrupted COIL object database. Note that LDA and our approach obtain at most $c-1$ features, where c is the number of classes. (a) 20 classes. (b) 40 classes. (c) 60 classes. (d) 80 classes. (e) 20 classes. (f) 40 classes. (g) 60 classes. (h) 80 classes.

TABLE I

AVERAGE RECOGNITION RATES (%) WITH STANDARD DEVIATIONS OF ALL COMPARED METHODS ON COIL OBJECT DATABASE

Methods	Original Images					
	20 objects	40 objects	60 objects	80 objects	100 objects	Average
PCA [1]	86.42±1.11	83.75±1.12	81.01±0.92	80.53±0.78	82.75±0.59	82.89±2.36
LDA [2]	81.83±2.03	77.08±1.36	66.96±1.52	59.34±1.22	52.29±0.30	67.50±12.19
NPE [4]	82.24±2.25	76.01±1.04	63.22±1.36	52.18±1.44	30.73±1.31	60.88±20.47
LSDA [5]	82.79±1.70	75.01±1.14	62.85±1.41	51.69±2.05	26.77±1.05	59.82±21.94
RPCA+LDA	83.26±1.52	78.39±1.15	68.93±0.86	60.73±0.68	56.44±0.73	69.55±11.36
SVM [54]	86.52±1.51	86.73±1.40	82.30±0.84	77.42±1.12	81.91±0.88	82.98±3.84
FDDL [55]	87.29±1.78	85.18±1.10	83.52±0.67	78.47±1.09	76.23±1.46	82.14±4.64
LatLRR [9]	88.98±0.85	88.45±0.64	86.36±0.52	84.67±0.79	82.64±0.60	86.22±2.64
DLRD [25]	89.58±1.04	86.79±0.94	82.60±1.06	81.10±0.58	79.92±0.93	84.00±4.06
Ours	**92.03**±1.21	**92.51**±0.65	**90.82**±0.43	**88.75**±0.71	**85.12**±0.33	**89.85**±3.01
Methods	10% Corrupted Images					
	20 objects	40 objects	60 objects	80 objects	100 objects	Average
PCA [1]	71.43±1.12	70.22±1.56	69.80±0.65	67.84±0.83	65.68±0.76	68.99±2.26
LDA [2]	47.77±3.06	45.89±1.12	36.42±1.12	27.13±0.95	16.79±0.34	34.80±13.01
NPE [4]	50.75±2.37	45.29±1.24	31.49±1.71	18.49±0.98	14.25±0.24	32.05±16.02
LSDA [5]	53.36±2.79	43.27±1.94	31.61±2.79	18.61±1.50	13.10±0.22	31.99±16.73
RPCA+LDA	49.35±1.55	53.26±1.84	44.18±2.65	29.92±0.96	23.55±0.46	40.05±12.78
SVM [54]	80.44±1.76	75.98±1.15	72.27±0.46	67.38±0.81	65.00±0.78	72.21±6.27
FDDL [55]	70.17±1.13	60.46±0.79	49.88±0.49	41.52±0.71	40.24±0.54	52.45±12.78
LatLRR [9]	81.38±1.25	81.93±0.92	80.97±0.45	77.15±0.72	73.47±0.62	78.98±3.61
DLRD [25]	82.96±1.81	80.77±0.95	76.93±1.25	74.03±0.74	73.82±0.77	77.70±4.07
Ours	**86.45**±1.12	**82.03**±1.31	**82.05**±0.87	**79.83**±0.62	**74.95**±0.65	**81.06**±4.18

derived from our approach is robust to pixel corruption. Table I shows the average recognition rates with standard deviations of all compared methods. It can be observed from Table I that the recognition rates of our approach vary slightly when the number of classes increases from 20 to 100.

The total average results are also summarized in Table I. We can see that our approach and LatLRR have lower deviations than other methods, which demonstrates good scalability. When the images are corrupted, all traditional subspace methods have difficulty obtaining reasonable results. However, three low-rank modeling-based methods achieve remarkable performance. In most cases, our SRRS approach achieves the best recognition results. Moreover, we utilize other levels of corruption, such as 20%, 30%, 40%, and 50% on COIL-20 database, and report the results in Fig. 4. It shows that our SRRS approach consistently outperforms other methods.

We also performed a significance test, McNemar's test, for the results shown in Table I, in order to demonstrate the statistical significance of our approach compared with several of the most representative state-of-the-art methods. We use a significance level of 0.05. In another word, the performance difference between two methods is statistically significant, if the estimated p-value is lower than 0.05. Table II shows the

TABLE II

p-VALUE BETWEEN SRRS AND OTHER METHODS ON THE COIL OBJECT DATABASE. THE ASTERISK * INDICATES THAT THE DIFFERENCE BETWEEN METHOD A AND METHOD B IS STATISTICALLY SIGNIFICANT WHEN $p = 0.05$

Method A vs. B	Original Images				Corrupted Images			
	20 objects	40 objects	60 objects	80 objects	20 objects	40 objects	60 objects	80 objects
Ours vs. PCA [1]	$1.0\times10^{-6*}$	$3.2\times10^{-8*}$	$2.3\times10^{-9*}$	$2.5\times10^{-7*}$	$1.2\times10^{-8*}$	$2.3\times10^{-10*}$	$3.5\times10^{-10*}$	$3.1\times10^{-10*}$
Ours vs. LDA [2]	$1.5\times10^{-7*}$	$5.5\times10^{-9*}$	$4.7\times10^{-9*}$	$2.4\times10^{-8*}$	$1.1\times10^{-9*}$	$1.1\times10^{-10*}$	$5.4\times10^{-10*}$	$1.7\times10^{-10*}$
Ours vs. NPE [4]	$2.3\times10^{-8*}$	$2.7\times10^{-8*}$	$1.6\times10^{-8*}$	$5.2\times10^{-9*}$	$1.6\times10^{-9*}$	$4.9\times10^{-10*}$	$2.9\times10^{-11*}$	$5.1\times10^{-10*}$
Ours vs. LSDA [5]	$1.3\times10^{-8*}$	$2.0\times10^{-8*}$	$2.5\times10^{-8*}$	$2.3\times10^{-8*}$	$1.2\times10^{-9*}$	$4.1\times10^{-10*}$	$3.8\times10^{-10*}$	$2.0\times10^{-10*}$
Ours vs. RPCA+LDA	$3.4\times10^{-8*}$	$3.2\times10^{-8*}$	$5.0\times10^{-8*}$	$2.9\times10^{-8*}$	$4.7\times10^{-9*}$	$2.5\times10^{-10*}$	$1.3\times10^{-10*}$	$1.1\times10^{-10*}$
Ours vs. SVM [54]	$3.1\times10^{-5*}$	$5.8\times10^{-6*}$	$1.2\times10^{-5*}$	$3.3\times10^{-8*}$	$1.0\times10^{-9*}$	$2.2\times10^{-9*}$	$1.8\times10^{-9*}$	$1.2\times10^{-10*}$
Ours vs. FDDL [55]	$2.7\times10^{-6*}$	$4.6\times10^{-7*}$	$3.3\times10^{-8*}$	$1.2\times10^{-7*}$	$2.1\times10^{-9*}$	$2.9\times10^{-10*}$	$1.4\times10^{-9*}$	$1.7\times10^{-9*}$
Ours vs. LatLRR [9]	$3.5\times10^{-5*}$	$2.1\times10^{-6*}$	$4.9\times10^{-5*}$	$3.5\times10^{-2*}$	0.0132*	0.0511	0.0920	0.0283*
Ours vs. DLRD [25]	0.0279*	$7.0\times10^{-4*}$	$2.1\times10^{-8*}$	$1.4\times10^{-8*}$	0.1325	$3.1\times10^{-5*}$	$2.4\times10^{-5*}$	$1.3\times10^{-4*}$

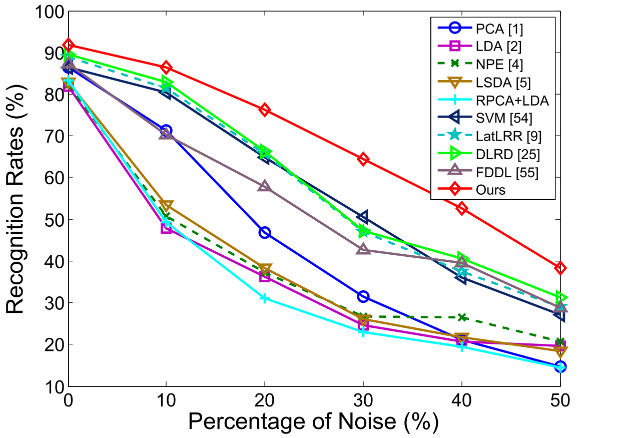
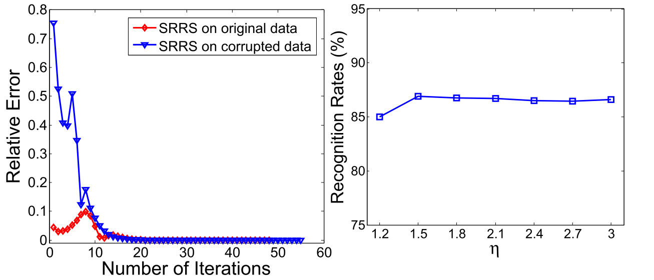


Fig. 4. Average recognition rates of all compared methods on COIL database with different levels of noise.



(a) (b)

Fig. 5. Properties of our approach on ALOI data set. (a) Convergence curve ($\rho = 1.3$, $\mu = 0.1$, and $\epsilon = 10^{-8}$). (b) Recognition rates of SRRS with different values of η.

p-values of comparing SRRS with other methods. From this table, the following conclusions can be reached.

1) The performance differences between our approach and the methods (PCA, LDA, NPE, LSDA, RPCA + LDA, SVM, and FDDL) are statistically significant in all cases.
2) On the original data set, the performance differences between our approach and DLRD/LatLRR are statistically significant.
3) On the corrupted data set, the performance differences between our approach and DLRD/LatLRR are not statistically significant. The reason is that DLRD and LatLRR are also able to handle the noisy data. But our approach achieves higher recognition rates than them.

The ALOI data set contains 1000 general object categories taken at different viewing angles. There are 72 equally spaced views in each category. In our experiments, we select the first 300 objects from this data set. All the images are converted to grayscale and resized to the size of 36 × 48. We also add 10% pixel corruption on the original images to evaluate the performance of different methods. Some examples of corrupted images in the ALOI data set can be found in Fig. 9.

Ten images of each object are randomly selected as training samples, and the others as test samples. This random selection process was repeated 20 times. Fig. 5(a) shows the convergence curves of our approach on the original data and the corrupted data. It shows that the relative error on corrupted data (10% noise) is larger than that on the original data. But, in both cases, our approach converges very well after ten iterations. The relative error is calculated by $\|X - AZ - E\|_F / \|X\|_F$. Fig. 5(b) shows the recognition rates of SRRS when parameter η is selected from the range [03]. We observe that SRRS is not very sensitive to the choice of η when $\eta > 1$. Furthermore, we set η to 1.5 to achieve the best recognition performance. Table III shows the average recognition rates with standard deviations for each compared method. It shows that SVM, FDDL, and two low-rank methods obtain better performance than the traditional subspace methods, and our approach outperforms all these methods on the original data set and the corrupted data set. In addition, by comparing Tables I and III, we can observe that, for the data set with a large number of classes, the classification task becomes more difficult when data are corrupted.

TABLE III

AVERAGE RECOGNITION RATES (%) ON ALOI OBJECT DATABASE

Methods	Original images	10% Corruption
PCA [1]	84.10±0.51	22.99±0.29
LDA [2]	83.46±0.38	23.97±0.37
NPE [4]	84.50±0.43	24.83±0.25
LSDA [5]	83.96±0.35	24.05±0.31
RPCA+LDA	84.62±0.48	27.58±0.41
SVM [54]	84.61±0.45	30.29±0.33
FDDL [55]	84.77±0.72	23.03±0.42
LatLRR [9]	84.97±0.53	59.35±0.48
DLRD [25]	85.53±0.55	58.66±0.39
Ours	**87.91**±0.34	**64.62**±0.32

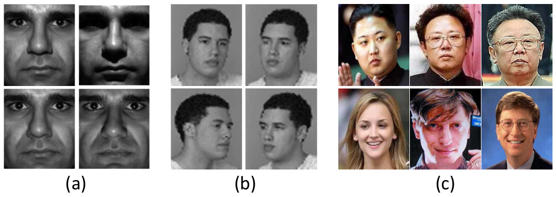


Fig. 6. Sample images in (a) YaleB, (b) FERET, and (c) KinFace data sets.

TABLE IV
AVERAGE RECOGNITION RATES (%) OF ALL COMPARED METHODS ON YALEB AND FERET FACE DATABASES

Methods	YaleB	FERET
PCA [1]	72.57±0.58	84.00±2.11
LDA [2]	89.09±0.91	77.63±2.22
NPE [4]	86.01±1.37	71.67±1.95
LSDA [5]	92.94±0.88	73.27±3.01
RPCA+LDA	91.29±1.16	79.03±2.63
SVM [54]	94.93±0.75	88.03±2.04
FDDL [55]	95.10±1.31	86.00±2.51
LatLRR [9]	88.76±1.26	84.27±2.19
DLRD [25]	93.56±1.25	83.33±2.40
Ours	**97.75**±0.58	**89.84**±2.01

B. Face Recognition With Illumination and Pose Variation

We also evaluate our approach on the Extended YaleB [51] and the FERET [52] face databases. The YaleB face data set consists of 2414 frontal face images of 38 classes, and each of them contains about 64 images. Fig. 6(a) shows the examples from the YaleB data set. We crop and resize the images to the size of 28 × 32, and normalize the pixel values to [0, 1].

As suggested in [9], we randomly select 30 images per class to construct the training set, and test set contains the rest of the images. This random selection procedure is repeated 20 times, and we show the average recognition rates in Table IV. It can be observed that supervised methods perform much better than the unsupervised method PCA. The reason is that PCA has a high sensitivity to illumination effects contained in this database. Due to the low-rankness property, the unsupervised method LatLRR greatly improves the recognition rate of PCA. In addition, the supervised low-rank method DLRD obtains higher recognition rate than LatLRR. By incorporating supervised information and low-rankness property, our approach can achieve an average recognition rate of 97.17% and outperform all the other methods, which implies that our approach is robust to variation illumination.

To evaluate the robustness to noise of the different methods, we randomly choose a percentage (from 10% to 50%) pixels and replace their values by random numbers that are uniformly distributed on [0, 1]. Fig. 7 shows that, in noisy scenarios, low-rank modeling-based methods (LatLRR, DLRD, and our approach) consistently obtain better performance than other methods. In particular, our SRRS approach can get the best performance.

The FERET database contains 2200 face images collected from 200 subjects, and each subject has 11 images. These images were captured under various poses and expressions. In this experiment, we randomly select the images from 50 individuals. Fig. 6(b) shows the images of one individual that show large pose variations. The original size of each image is 384 × 256. We cropped and resized them to the size of 30 × 25.

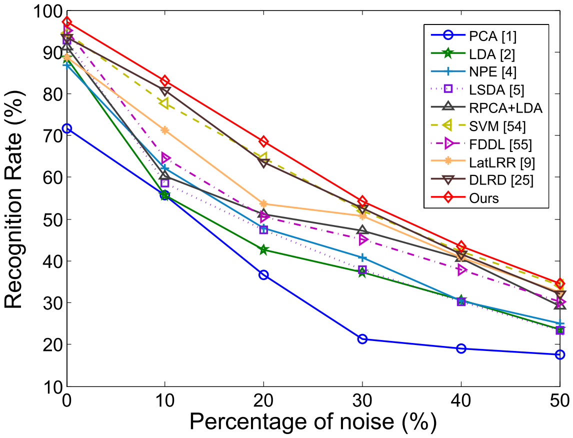


Fig. 7. Average recognition rates of all compared methods on YaleB database with different levels of noise.

We randomly select five images of each individual as training samples, and the remaining samples are regarded as test samples. Table IV lists the average recognition rates of all compared methods over 20 runs. It reflects that our approach can improve the recognition results over the existing methods. Interestingly, the PCA can outperform some supervised subspace methods on this database. A likely reason for this is that large pose changes of one individual produce large intraclass variations, which highly influence the performance of supervised methods.

C. Face Recognition With Occlusions

The AR face database contains over 4000 facial images collected from 126 subjects. For each subject, there are 26 frontal face images, taken under different illuminations, expressions, and facial occlusions in two separate sessions. In our experiments, we strictly follow the experimental settings in [26], and conducted the following three experiments.

1) Sunglasses: Some face images contain the occlusion of sunglasses, which are considered as corrupted samples. To construct the training set, we choose seven neutral images and one randomly selected image with sunglasses from each subject (session 1). The test set contains the remaining neutral images (session 2) and the rest of the images with sunglasses (two images from sessions 1 and 3 images from session 2). Thus, for each individual, there are 8 training images and 12 test images. The sunglasses cover about 20% of the face image.

2) Scarf: We utilize the corrupted training images due to the occlusion of scarf. Using a similar training/test setting as above, we have 8 training images and 12 test images for each individual. The scarf covers about 40% of the face image.

3) Sunglasses + Scarf: Moreover, we consider the case where images contain both sunglasses and scarf. We select all the seven neutral images and two corrupted images (one with sunglasses and the other with scarf) at session 1 for training. For each individual, there are 17 test images in total.

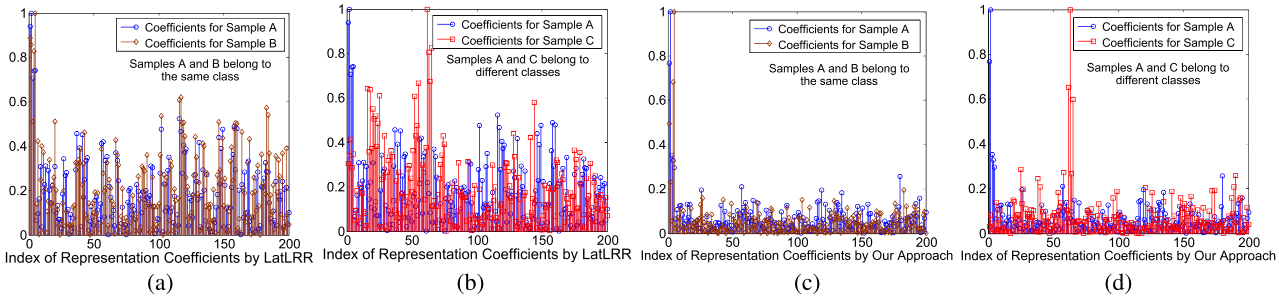


Fig. 8. Visualization of LRR coefficients of two pairs of samples on FERET database. (a) LatLRR: same class. (b) LatLRR: different classes. (c) Ours: same class. (d) Ours: different classes.

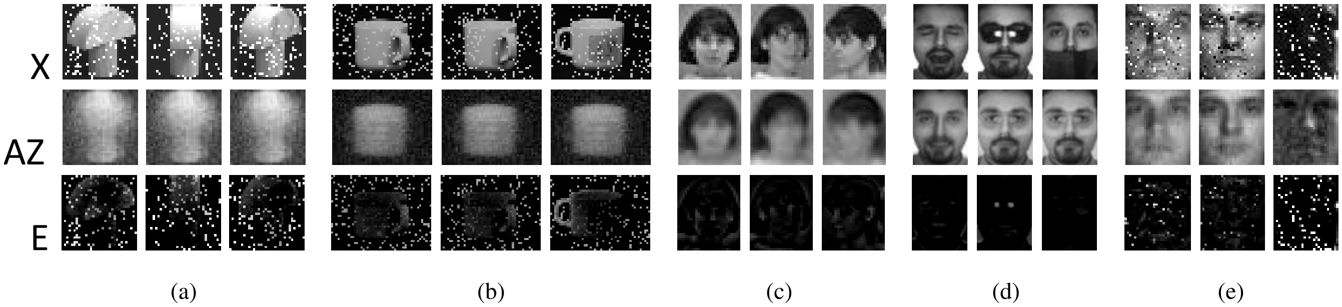


Fig. 9. Visualization of corrupted images (X), recovered images (AZ), and noise (E) on five object and face databases. (a) COIL. (b) ALOI. (c) FERET. (d) AR. (e) YaleB.

TABLE V
RECOGNITION RATES (%) OF ALL COMPARED METHODS ON AR FACE DATABASES

Methods	Sunglasses	Scarf	Sunglasses+Scarf
PCA [1]	51.75	48.83	42.00
LDA [2]	82.25	81.75	79.82
NPE [4]	83.75	82.58	80.35
LSDA [5]	83.66	81.83	79.12
SVM [54]	78.37	65.33	67.71
LatLRR [9]	76.52	75.24	76.11
DLRD [25]	85.26	83.01	81.56
LRDL [26]	**87.21**	83.96	82.15
Ours	85.84	**86.98**	**86.23**

Table V shows the recognition rates of compared methods in three different scenarios. We can observe that LRDL obtains the best result in the sunglasses case, and our approach obtains the best results in the scarf case and the mixed case. Fig. 9(d) shows that our approach can correctly recover the clean images from the occluded face images. Therefore, we can train robust classifiers from the recovered image set AZ.

D. Kinship Verification

Kinship verification is a recently investigated research topic, which aims at determining kin relationships from photos. It is still a very challenging task due to large variations in different human faces. We also evaluate the performance of our approach and related methods on kinship verification. We conduct the kinship verification experiments on the UB KinFace database Version 2 [53], [56]. This database contains 600 face images that can be separated into 200 groups, and each group consists of children, young parents, and old parents. Fig. 6(c) shows example images in the KinFace database, in which three columns (from left to right) represent the images of children, young parents, and old parents, respectively. Given two images of faces, our task is to determine whether they are an accurate child–parent pair.

As suggested in [56], we employ the difference vectors between the child and the parent as the features rather than directly compare children with their parents. In particular, in the experiments for children and old parents, we build 200 true child–old parent pairs and 200 false child–old parent pairs. The experiments for children and young parents are carried out in a similar manner. Then, we conduct fivefold cross validation for this verification problem. At each round, 160 true pairs and 160 false pairs are used for training, and the rest is used for testing. Average verification rates are reported in Table VI. Our approach outperforms all the other methods. In this binary classification problem, some traditional supervised methods perform very poorly.

E. Discussion

The experimental results show that, compared with the traditional subspace learning methods, our approach is robust to noise and large variations. The reason is that the low-rank property helps us obtain a better estimate of the underlying distribution of samples from the recovered images, and then, our approach learns a robust and discriminative subspace. The resulting performance is better than the compared low-rank modeling and dictionary learning methods.

Fig. 8 shows why our approach performs so well by visualizing LRR coefficients of LatLRR and our approach. In particular, we show the coefficients for representing

TABLE VI

Verification Rates (%) on UB KinFace Database (Fivefold Cross Validation). C Versus Y and C Versus O Denote Child–Young Parent Verification and Child–Old Parent Verification, Respectively

Methods	C vs. Y	C vs. O
PCA [1]	57.25±2.59	56.75±2.05
LDA [2]	47.58±5.36	49.25±7.27
NPE [4]	55.25±3.01	55.75±3.60
LSDA [5]	56.75±3.24	57.00±2.88
RPCA+LDA	56.25±5.08	52.00±4.56
SVM [54]	51.00±2.05	49.25±1.43
FDDL [55]	48.00±2.21	51.46±3.51
LatLRR [9]	55.00±3.53	52.50±2.76
DLRD [25]	53.63±3.19	54.27±2.88
Ours	**61.79**±2.13	**62.15**±2.23

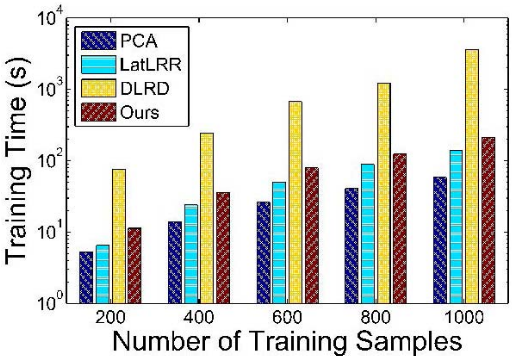


Fig. 10. Training time (seconds) with different numbers of training samples on COIL object database.

two pairs of samples. One pair, samples A and B, is selected from the same class, while the other pair, samples A and C, is selected from different classes. Fig. 8(a) and (b) shows that, in LatLRR, samples from the same class contribute more in the representation, as the coefficients within the same class are a little larger than those in other classes. In some sense, LatLRR could discover the subspace membership of samples. Compared with LatLRR, Fig. 8(c) and (d) shows that the coefficients of the same class are higher than others, which implies our approach can clearly reveal the subspace structure. Since we incorporate supervised regularization in our model, the LRRs as well as the resulting subspace learnt by our approach should be more discriminative than that of the LatLRR.

Furthermore, Fig. 9 shows the corrupted images, the recovered images, and the noisy part on five object and face databases. It shows that, although training images (i.e., X) have large pose variations and corruptions, the recovered images AZ are very similar to each other, which helps us learn a robust subspace for classification.

We evaluate the computational cost of different methods when increasing the sample size. Taking COIL database as an example, the training times are shown in Fig. 10. Since PCA, LDA, NPE, and LSDA have similar computational complexity, and FDDL has a similar complexity to DLRD, we only compare against LatLRR, PCA, and DLRD. In Fig. 10, we observe that the linear subspace method PCA has the lowest training time. Our approach and LatLRR have similar training time, which is much less than the time cost of DLRD. Moreover, the test time of our approach is even less than that of PCA, due to the fact that our approach can achieve the best recognition rates with only a few features.

V. Conclusion

In this paper, a novel LSL approach, SRRS, is proposed for feature extraction and classification. The proposed approach iteratively learns robust subspaces from a low-rank learning model, and naturally incorporates discriminative information. The convexity of the supervised regularization term has been theoretically proved. The experimental results on six benchmark data sets demonstrate the effectiveness of our approach compared with the state-of-the-art subspace methods and low-rank learning methods. Moreover, when the data contain considerable noise or variations, our approach can improve the classification performance.

In our future work, we will develop a divide-and-conquer version of SRRS approach to make it scalable to larger data sets, and we would also like to design the dictionary learning algorithms to further enhance the classification performance.

References

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

[2] 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.

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

[4] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1208–1213.

[5] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, Jan. 2007, pp. 708–713.

[6] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

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

[8] T. Mu, J. Goulermas, J. Tsujii, and S. Ananiadou, "Proximity-based frameworks for generating embeddings from multi-output data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2216–2232, Nov. 2012.

[9] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1615–1622.

[10] 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.

[11] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 625–632.

[12] 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.

[13] L. Zhang, M. Yang, Z. Feng, and D. Zhang, "On the dimensionality reduction for sparse representation based face recognition," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 1237–1240.

[14] L. Zhang, P. Zhu, Q. Hu, and D. Zhang, "A linear subspace learning approach via sparse coding," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 755–761.

[15] Z. Lai, Y. Li, M. Wan, and Z. Jin, "Local sparse representation projections for face recognition," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2231–2239, 2013.

[16] 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.

[17] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2011, Art. ID 11.

[18] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[19] S. Li and Y. Fu, "Learning balanced and unbalanced graphs via low-rank coding," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1274–1287, May 2015.

[20] S. Li and Y. Fu, "Low-rank coding with *b*-matching constraint for semi-supervised classification," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Aug. 2013, pp. 1472–1478.

[21] S. Li, M. Shao, and Y. Fu, "Multi-view low-rank analysis for outlier detection," in *Proc. SIAM Int. Conf. Data Mining*, May 2015, pp. 748–756.

[22] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2168–2175.

[23] M. Shao, C. Castillo, Z. Gu, and Y. Fu, "Low-rank transfer subspace learning," in *Proc. 12th IEEE Int. Conf. Data Mining*, Dec. 2012, pp. 1104–1109.

[24] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 74–93, 2014.

[25] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2586–2593.

[26] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 676–683.

[27] L. Li, S. Li, and Y. Fu, "Learning low-rank and discriminative dictionary for image classification," *Image Vis. Comput.*, vol. 32, no. 10, pp. 814–823, 2014.

[28] Y. Pan, H. Lai, C. Liu, and S. Yan, "A divide-and-conquer method for scalable low-rank latent matrix pursuit," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 524–531.

[29] A. Talwalkar, L. Mackey, Y. Mu, S.-F. Chang, and M. I. Jordan, "Distributed low-rank subspace segmentation," in *Proc. 14th IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3543–3550.

[30] Y.-X. Wang, H. Xu, and C. Leng, "Provable subspace clustering: When LRR meets SSC," in *Proc. 27th Annu. Conf. Adv. Neural Inf. Process. Syst.*, 2013, pp. 64–72.

[31] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. 25th Annu. Conf. Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[32] S. Li and Y. Fu, "Robust subspace discovery through supervised low-rank constraints," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 163–171.

[33] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[34] M. Zhu and A. M. Martínez, "Subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1274–1286, Aug. 2006.

[35] F. R. Bach, "Consistency of trace norm minimization," *J. Mach. Learn. Res.*, vol. 9, pp. 1019–1048, Jun. 2008.

[36] Y. Deng, Q. Dai, R. Liu, Z. Zhang, and S. Hu, "Low-rank structure learning via nonconvex heuristic recovery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 3, pp. 383–396, Mar. 2013.

[37] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "TILT: Transform invariant low-rank textures," *Int. J. Comput. Vis.*, vol. 99, no. 1, pp. 1–24, 2012.

[38] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[39] C. Chen, J. Cai, W. Lin, and G. Shi, "Surveillance video coding via low-rank and sparse decomposition," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 713–716.

[40] G. Liu and S. Yan, "Active subspace: Toward scalable low-rank learning," *Neural Comput.*, vol. 24, no. 12, pp. 3371–3394, 2012.

[41] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2618–2625.

[42] N. Zhang and J. Yang, "Low-rank representation based discriminative projection for robust feature extraction," *Neurocomputing*, vol. 111, pp. 13–20, Jul. 2013.

[43] Z. Zheng *et al.*, "Low-rank matrix recovery with discriminant regularization," in *Proc. 17th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, Apr. 2013, pp. 437–448.

[44] Y.-F. Guo, S.-J. Li, J.-Y. Yang, T.-T. Shu, and L.-D. Wu, "A generalized Foley–Sammon transform based on generalized Fisher discriminant criterion and its application to face recognition," *Pattern Recognit. Lett.*, vol. 24, nos. 1–3, pp. 147–158, 2003.

[45] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Mar. 2010.

[46] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[47] J. K. Merikoski and R. Kumar, "Inequalities for spreads of matrix sums and products," *Appl. Math. E-Notes*, vol. 4, pp. 150–159, Feb. 2014.

[48] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation $AX + XB = C$ [F4]," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, 1972.

[49] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Columbia Univ. Comput. Sci., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

[50] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 103–112, 2005.

[51] K.-C. Lee, J. Ho, and D. 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.

[52] P. J. Phillips, H. Moon, S. A. Rozvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[53] S. Xia, M. Shao, and Y. Fu, "Kinship verification through transfer learning," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2539–2544.

[54] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York, NY, USA: Springer-Verlag, 2000.

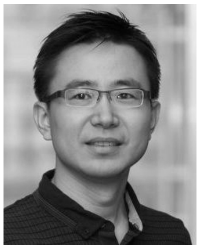
[55] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 543–550.

[56] S. Xia, M. Shao, J. Luo, and Y. Fu, "Understanding kin relationships in a photo," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1046–1056, Aug. 2012.

Sheng Li (S'11) received the B.Eng. degree in computer science and engineering and the M.Eng. degree in information security from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2010 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA.

He was a Data Scientist Intern with Adobe Research, San Jose, CA, USA, in 2014 and 2015. His current research interests include low-rank matrix recovery, data mining, and machine learning.

Mr. Li was a recipient of the best paper award at the 2014 SIAM International Conference on Data Mining (SDM) and the Best Student Paper Honorable Mention Award at the 2013 IEEE International Conference on Automatic Face and Gesture Recognition. He has received the ACM SIGIR Student Travel Award for the International Conference on Information and Knowledge Management in 2015, and the NSF Student Travel Awards for SDM and the International Conference on Data Mining in 2014. He serves as a Reviewer for several IEEE TRANSACTIONS, and a Program Committee Member of the International Joint Conference on Artificial Intelligence.



Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, Xi'an, China, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA.

He is currently an Interdisciplinary Faculty Member with the College of Engineering and the College of Computer and Information Science, Northeastern University, Boston, MA, USA. His current research interests include interdisciplinary research in machine learning and computational intelligence, social media analytics, human–computer interaction, and cyber-physical systems.

Dr. Fu is a Lifetime Member of the Association for Computing Machinery, the Association for the Advancement of Artificial Intelligence, the International Society for Optics and Photonics, and the Institute of Mathematical Statistics, and a member of the International Neural Network Society. He was a Beckman Graduate Fellow from 2007 to 2008. He was a recipient of five best paper awards (SIAM International Conference on Data Mining in 2014, the IEEE International Conference on Automatic Face and Gesture Recognition in 2013, the IEEE ICDM Large Scale Visual Analytics Workshop in 2011, the IAPR International Conference on Frontiers in Handwriting Recognition in 2010, and the IEEE International Conference on Image Processing in 2007), three young investigator awards (the ONR Young Investigator Award in 2014, the ARO Young Investigator Award in 2014, and the INNS Young Investigator Award in 2014), the National Academy of Engineering U.S. Frontiers of Engineering in 2015, two service awards (the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Associate Editor in 2012 and the IEEE ICME Best Reviewer in 2011), the IC Post-Doctoral Research Fellowship Award in 2011, and the Google Faculty Research Award in 2010. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.