Learning Latent Low-Rank and Sparse Embedding for Robust Image Feature Extraction

Zhenwen Ren, Quansen Sun, Bin Wu, Xiaoqian Zhang, Wenzhu Yan

***Abstract*—To defy the curse of dimensionality, the inputs are always projected from the original high-dimensional space into the target low-dimension space for feature extraction. However, due to the existence of noise and outliers, the feature extraction task for corrupted data is still a challenging problem. Recently, a robust method called low rank embedding (LRE) was proposed. Despite the success of LRE in experimental studies, it also has many disadvantages: 1) The learned projection cannot quantitatively interpret the importance of features. 2) LRE does not perform data reconstruction so that the features may not be capable of holding the main energy of the original "clean" data. 3) LRE explicitly transforms error into the target space. 4) LRE is an unsupervised method, which is only suitable for unsupervised scenarios. To address these problems, in this paper, we propose a novel method to exploit the latent discriminative features. In particular, we first utilize an orthogonal matrix to hold the main energy of the original data. Next, we introduce an $\ell_{2,1}$-norm term to encourage the features to be more compact, discriminative and interpretable. Then, we enforce a columnwise $\ell_{2,1}$-norm constraint on an error component to resist noise. Finally, we integrate a classification loss term into the objective function to fit supervised scenarios. Our method performs better than several state-of-the-art methods in terms of effectiveness and robustness, as demonstrated on six publicly available datasets.**

***Index Terms*—subspace learning, feature extraction, low-rank embedding, $\ell_{2,1}$-norm, face recognition**

I. Introduction

HIGH-DIMENSIONAL features have been widely used in computer vision and pattern recognition tasks [1], such as speech recognition, gene expression, text document clustering, face recognition, and object classification. However, irrelevant and redundant features may 1) cause heavy computational complexity and memory consumption, 2) increase the possibility of over-fitting, and 3) greatly weaken the inherent structural relations between features, which ultimately degrade algorithmic performance. Thus, finding compact low-dimension representation features with good generalization and discriminative abilities is critical.

Z. Ren is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with School of National Defence Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China (e-mail: rzw@njust.edu.cn).

B. Wu is with the School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China (e-mail: wb@swust.edu.cn).

Q. Sun, W. Yan, and X. Zhang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: sunquansen@njust.edu.cn; ywznanj@163.com; zhxq0528@163.com).

Q. Sun is the corresponding author.

Manuscript received $* *$, 2019.

There are two classic feature extraction methods: principal component analysis (PCA) [2] and linear discriminant analysis (LDA) [3]. The former is unsupervised, and the latter is supervised [4]. Both of them are linear and assume a Gaussian distribution of the input data. Among them, PCA is probably the most well-known and widely used data-processing and dimension reduction technology. It seeks to find the main uncorrelated directions in which the data can capture the maximal variance. However, the performance of PCA deteriorates rapidly when data are corrupted. Later, many extended PCA-based methods were developed to overcome the influence of noise. For example, neighborhood preserving embedding (NPE) [5], locality preserving projection (LPP) [6], and sparsity preserving projections (SPP) [7]. LDA is another favorite supervised dimensionality reduction technology, which seeks for a discriminant projection by simultaneously maximizing the scatter between the class centers and minimizing the scatter within each class. However, LDA fails when the data have a non-Gaussian distribution or there is a small sample size. To overcome these and other limitations, subsequent LDA-based methods have been proposed, such as orthogonal LDA (OLDA) [8], manifold partition discriminant analysis (MPDA) [9], and discriminative locality alignment (DLA) [10].

Recently, representation-based feature learning methods have drawn substantial attention [11–15]. Sparse representation (SR) and low-rank representation (LRR) are two of the most famous benchmark methods. The first representation-based method may be sparse representation classification (SRC) [12]. SRC represents a test sample by seeking for a sparsest linear combination over an over-complete dictionary, which usually requires the collection of all training samples of all classes as the dictionary. It is reported that SRC achieves surprisingly high accuracy in face recognition, even with occlusion. However, SRC may depress the performance when the training data are corrupted by various noise and outliers, e.g., unreasonable illuminations, pose variations, occlusions, etc. Moreover, SRC is unable to expose the inherent global structure of the samples. To overcome these drawbacks, the low-rank representation (LRR) [13, 14] based feature learning method has been proposed. With the assumption that the samples from the same class should lay on an intrinsic low-dimensional subspace, the LRR can better expose the underlying global structure between samples to remove the adverse impacts of noise and outliers. Although the LRR has been successfully applied to image classification and subspace clustering tasks, it still has several limitations [16].

Largely inspired by both the classic feature extraction/selection methods and the representation-based feature

learning methods, many researchers have combined both to overcome their respective drawbacks. Thus, a series of novel and valuable methods [2, 15–27] have been produced that greatly improve the feature learning theory and practice. Accordingly, these methods have achieved great success in many fields. These methods include sparse principal component analysis (SPCA) [24], robust principal component analysis (RPCA) [2], collaborative representation-based classification (CRC) [28], discriminant sparse neighborhood preserving embedding (DSNPE) [22], robust latent subspace learning (RLSL) [25], sparse discriminant analysis (SDA) [26], robust sparse linear discriminant analysis (RSLDA) [23], latent low-rank representation (LatLRR) [20], supervised approximate low-rank projection matrix learning (SALPL) [29], low-rank representations for image classification (LRRC) [19], and low-rank approximation and structure learning for unsupervised feature selection (LRSL) [27]. These related methods are discussed in greater detail in the next section of this paper.

Low rank embedding (LRE) [30] is a recently proposed robust feature extraction technology, which can simultaneously search both the optimal low-rank representation and latent embedding subspace. Nonetheless, it has some shortcomings as well, which may degrade the performance. (1) The resulting discriminative projection Q cannot well interpret the importance of features, i.e., we do not know which of them are relatively more important. (2) LRE ignores the information loss of space mapping since it does not consider data reconstruction from the embedding subspace to the original space, i.e., the representation $Q^T X$ cannot preserve the primary energy of the original "clean" data. (3) LRE explicitly transforms the sparse error into the target space, which is not conducive to separating noise. (4) LRE is an unsupervised method, which is only suitable for the unsupervised scenario but not for the supervised scenario.

As far as we know, not just LRE but some SR and LRR-based methods also have these problems that may potentially degrade their performance. Motivated by LRE, to solve the above problems and obtain more discriminative features, in this paper, we propose a new method named latent low-rank and sparse embedding (LLRSE). In summary, our contributions are highlighted as follows:

1) We combine a data reconstruction term with some regularization terms to form a unified objective function. To do this, we first introduce an orthogonal reconstruction matrix, which can provide more freedom to guarantee that the main energy is held in the obtained low-dimensional embedding space and thus is appropriate for performing subsequent tasks. Second, we embed an $\ell_{2,1}$-norm regularization term of projection matrix Q into the objective function, which can simultaneously extract and select the most discriminative features while providing good interpretability for the importance of features. Third, by assuming that samples within the same class implicitly have a low-rank structure, we employ a low-rank matrix Z to capture the inherent global correlation structure of the data so that the corrupted data can be approximately recovered via the low-rank recovery. Finally, we utilize an error component E with a columnwise $\ell_{2,1}$-norm constraint to fit the error, which makes LLRSE robust to noise and outliers.

2) To ensure that the extracted features are optimal for classification purpose, we further integrate a classification loss function into the training process, which elegantly incorporates the label information and a linear classifier.

3) To efficiently solve the resulting optimization problem, we propose an efficient algorithm based on the alternating direction method of multipliers (ADMM) framework, which makes the objective function quickly converge.

4) Extensive experimental results demonstrate the good effectiveness, noise-resistibility, robustness, and reconstruction ability of the proposed method on a variety of complex classification tasks, especially with noise and corrupted observations.

The rest of this paper is structured as follows. In Section II, we briefly review for some closely related representation-based feature learning methods. In Sections III, we first propose the LLRSE model, and then provide the solution of the optimum model and present its computational complexity analysis to it. In the subsequent Section IV, the experimental settings and results are presented. The conclusion is shown in Section V.

II. Related Work

As the basic inspiration of the proposed method, we briefly review two categories of related representation-based methods: 1) ℓ_1, ℓ_2, and $\ell_{2,1}$ norm representation, and 2) low-rank representation.

A. ℓ_1, ℓ_2, and $\ell_{2,1}$ Norm Constrained Feature Learning

Inspired by the success of compressive sensing (CS) theory [31], SRC [12] assumes that a test sample $x \in \mathbb{R}^m$ will approximately lie in the linear span of an over-complete semantic dictionary $D = [D_1, \cdots, D_C] \in \mathbb{R}^{m \times r}$, where r is the number of atoms of D, and D_k is the k-th subdictionary. To code x, many sparse representation-based supervised feature learning methods can always be written as the following ℓ_1-norm optimization problem (Note: the ℓ_1-norm is used to replace the NP-hard ℓ_0-norm):

$$\min_{\alpha} \|x - D\alpha\|_2^2 + \beta\|\alpha\|_1 \tag{1}$$

where $\alpha = [\alpha_1^T, \alpha_2^T, \cdots, \alpha_C^T]^T$, α_k is the coding coefficient associated with the subdictionary D_k, and β is a regularization factor. Then, x is classified into the true class with the minimal coding residual. Although such a method has achieved great success in biological recognition, it requires the atoms in the semantic dictionary D to be well aligned, which is not always satisfied. The simplest strategy is to use the whole training set as the dictionary. Several extended methods have been proposed to address this issue [11, 17, 18, 25, 28]. For example, Wagner *et al.* [17] proposed an extended SRC method for undersampled face recognition under complex situations, such as variable illuminations, expressions, disguises, etc. Yang *et al.* [18] also extended SRC to deal with outliers, such as occlusions in face images. In addition, linear representation-based classification (LRC) [11] and collaborative representation-based classification (CRC) [28] have also achieved exciting performance for face recognition. Excepting ℓ_1-norm, ℓ_2-norm is also related to the discrimination of image feature [28]. When samples are not corrupted, the ℓ_2-norm is more suitable

for coding coefficients. However, a common drawback of these both ℓ_1-norm and ℓ_2-norm-based methods is that they cannot provide the consistent or jointly sparse representations.

Unlike the ℓ_1 norm and ℓ_2 norm, the $\ell_{2,1}$-norm has good row-sparsity property [23, 32, 33]. For example, Wen *et al.* [23] presented a feature selection method called robust sparse linear discriminant analysis (RSLDA) by imposing an $\ell_{2,1}$-norm regularization constraint on the projection matrix of linear discriminant analysis (LDA); the results demonstrate that the learned projection matrix has good interpretability for features and has a good rank property for selecting the most important features. Shi *et al.* [33] proposed an optimization mode named robust PCA via optimal mean (RPOM), which is constituted by combining the Schatten p-norm and $\ell_{2,1}$-norm.

B. Low-Rank Representation Based Feature Learning

Before we state the LRR-based feature learning methods, we first introduce the robust PCA (RPCA) [2] since some methods, including ours, are based on or related to it. RPCA is based on low-rank matrix recovery, which can efficiently expose the low-dimensional subspace structure by finding a low-rank component and a sparse error component to decompose the data matrix. The RPCA model is

$$\min_{A,E} \|A\|_* + \lambda\|E\|_1, \quad s.t.\ X = A + E \tag{2}$$

where $X \in \mathbb{R}^{m\times n}$ is the data matrix with n m-dimensional samples, $A \in \mathbb{R}^{m\times n}$ is a low-rank matrix, $E \in \mathbb{R}^{m\times n}$ is the sparse error component, and $\|\cdot\|_*$ is the nuclear norm that provides a good surrogate for the rank function. Although RPCA extracts a low-rank representation from the corrupted data, it is not a projection matrix.

Based on RPCA, the established low-rank representation (LRR) [13, 14, 27] is a widely used method that seeks the lowest-rank representation among all the training observations, and then presents each test sample as linear combinations of the atoms in a given dictionary $D \in \mathbb{R}^{m\times r}$, i.e.,

$$\min_{Z,E} \|Z\|_* + \lambda\|E\|_1, \quad s.t.\ X = DZ + E \tag{3}$$

where $Z \in \mathbb{R}^{r\times n}$ is the coefficient matrix. The LRR usually directly sets the data matrix itself as the dictionary (i.e., $D = X, r = n$) for simplicity. However, most real-life data are always corrupted and may degrade performance.

LatLRR [20, 29] is a recently proposed LRR-based features learning method for handling the case of insufficient sampling and/or grossly corrupted data. It is based on the premise that unobserved samples can represent observed samples. The LatLRR model is

$$\min_{Z,E} \|Z\|_* + \lambda\|E\|_1, \quad s.t.\ X = [X_O, X_H]Z + E \tag{4}$$

where X_O is the observed samples and X_H is the unobserved hidden samples. LatLRR uses two low-rank matrixes $Z \in \mathbb{R}^{n\times n}$ and $L \in \mathbb{R}^{m\times m}$ to learn the principal features XZ and salient features LX. Thus, the LatLRR is reformulated as

$$\min_{Z,L,E} \|Z\|_* + \|L\|_* + \lambda\|E\|_1, s.t.\ X = XZ + LX + E \tag{5}$$

However, the solution of the LatLRR is not unique, and such nonuniqueness makes the recognition performance of the LatLRR unpredictable [27]. Beyond that, Fang et al. [29] argued that the LatLRR has three disadvantages and proposed a supervised approximate low-rank projection matrix learning (SALPL) [29] method to address these disadvantages.

In LRE [30], the clean data in the embedding space can be approximately reconstructed by a low-rank matrix Z and an orthogonal transformation matrix $Q \in \mathbb{R}^{n\times d}$ via

$$\min_{Z,Q} \|Z\|_* + \lambda\|Q^TX - Q^TXZ\|_{2,1},\ s.t.\ Q^TQ = I. \tag{6}$$

From problem (6), we note that the reconstructive capacity is only measured by the term $\|Q^TX - Q^TXZ\|_{2,1}$ in the target embedding space, but it does not consider the information loss when the data are transformed from the original high-dimension space to the target low-dimension space. Moreover, LRE has some other disadvantages that were mentioned above; thus, the performance of LRE cannot be optimal.

III. The Proposed Method

In this section, a new robust image feature extraction method called LLRSE is first proposed. To solve LLRSE effectively, we develop an efficient ADMM-based optimization algorithm. The computational complexity is also presented.

A. Problem Formulation

As previously discussed, LRE [30] has some flaws, which make it disadvantageous for further improving performance. Since the data that are collected in real-world applications are naturally corrupted by noise and outliers [34] (e.g., illumination changes, pose variations, occlusion, etc.), we derive the following formulation:

$$\min_{Z,E,Q} \|Z\|_* + \lambda\|E\|_{2,1}, s.t.\ Q^TX = Q^TXZ + E, Q^TQ = I \tag{7}$$

where $X \in \mathbb{R}^{m\times n}$ is the data matrix, $E \in \mathbb{R}^{m\times n}$ denotes the error component, $Q \in \mathbb{R}^{m\times d}$ is the projection matrix, $Z \in \mathbb{R}^{n\times n}$ is the low-rank matrix, and $\lambda > 0$ is a regularization parameter. Note that the $\|\cdot\|_{2,1}$-norm is employed to instantiate $\|E\|_\ell$. The main reasons are as follows: When the noise is Gaussian noise, $\|E\|_\ell = \|E\|_F^2$, where $\|\cdot\|_F$ refers to the Frobenius norm. When the noise is due to entrywise corruptions, $\|E\|_\ell = \|E\|_1$, where $\|\cdot\|_1$ refers to the ℓ_1-norm. When the noise is due to sample-specific corruptions and outliers, $\|E\|_\ell = \|E\|_{2,1}$, where $\|E\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m} E_{ij}^2}$ refers to the column $\ell_{2,1}$-norm. Among them, $\|E\|_{2,1}$ can better characterize errors such as corruptions and outliers.

From problem (7), we know that the noise component of X in the original space is transformed into the target subspace (i.e., $Q^TX = Q^TXZ + E$), which is not conducive to separating the noise. Therefore, we explicitly avoid transforming the noise into the target space, but considering E in the original space. Moreover, a matrix $R \in \mathbb{R}^{m\times d}$ is introduced for performing data reconstruction, which gives Q more freedom to learn features. In this way, the embedding representation Q^TX will be more likely to preserve the primary energy of the "clean" data. Thus, we have the following objective function:

$$\begin{aligned} &\min_{Z,E,R,Q} \|Z\|_* + \lambda\|E\|_{2,1}, \\ &s.t.\ X = RQ^TX + E, Q^TX = Q^TXZ, Q^TQ = I. \end{aligned} \tag{8}$$

The above objective function implies a low-rank embedding component, i.e., $\|Q^TX-Q^TXZ\|_{2,1}$, which is latent low-rank embedding. Then, mathematically, (8) can be reduced to

$$\begin{aligned}&\min_{Z,E,R,Q}\|Z\|_*+\lambda\|E\|_{2,1},\\&s.t.\ X=RQ^TXZ+E, Q^TQ=I.\end{aligned}\tag{9}$$

Suppose that q_{ij} is the (i,j)-th element of the learned projection matrix $Q\in\mathbb{R}^{m\times d}$. For any sample x_i, the projected sample y is $y_i=Q^Tx_i$. More precisely, the relationship between x_i and y_i that is established through Q is

$$(y|Q^Tx)=\left[\begin{array}{c|cccc}y_{i1}&q_{11}x_{i1}&q_{21}x_{i2}&\cdots&q_{m1}x_{im}\\y_{i2}&q_{12}x_{i1}&q_{22}x_{i2}&\cdots&q_{m2}x_{im}\\\cdots&\cdots&\cdots&\cdots&\cdots\\y_{id}&q_{1d}x_{i1}&q_{2d}x_{i2}&\cdots&q_{md}x_{im}\end{array}\right]\tag{10}$$

where $(y|Q^Tx)$ is the augmented matrix of y and Q^Tx.

By forming Q with the $\ell_{2,0}$-norm constraint, the obtained projection matrix will be row sparse, i.e., most of unimportant or redundant features' corresponding rows are equal to zero. However, it is hard to efficiently solve the $\ell_{2,0}$-norm minimization problem since it is NP-hard. A natural alternative is to use the $\ell_{2,1}$-norm to better approximate the $\ell_{2,0}$-norm [30, 35]. In addition, an orthogonal constraint $R^TR=I$ is incorporated into our model to avoid a trivial solution. Consequently, we can obtain a more accurate Q and R by introducing an orthogonal constraint to R and removing the orthogonal constraint on Q. Hence, we have

$$\begin{aligned}&\min_{Z,E,R,Q}\lambda_1\|Q\|_{2,1}+\lambda_2\|E\|_{2,1}+\|Z\|_*,\\&s.t.\ X=RQ^TXZ+E, R^TR=I\end{aligned}\tag{11}$$

where $\lambda_1>0$ balances the row-sparsity of the feature projection Q, $\lambda_2>0$ balances the error component E, and $\|Q\|_{2,1}=\sum_{i=1}^m\sqrt{\sum_{j=1}^d Q_{ij}^2}$.

Based on the known label information, we further introduce a discriminant functional term to integrate the label information into the feature learning process. Consequently, the objective function can be reformulated as

$$\begin{aligned}&\min_{Z,E,R,Q}\lambda_1\|Q\|_{2,1}+\lambda_2\|E\|_{2,1}+\lambda_3\|Z\|_*+\Psi(X,Q)\\&s.t.\ X=RQ^TXZ+E, R^TR=I\end{aligned}\tag{12}$$

where $\lambda_3>0$ balances the low-rank representation term Z, and Ψ is the additional discriminate function that utilizes the known label information to learn discriminative features.

To learn a projection matrix with a strong discrimination capability for classification purpose, we construct a zero-one label matrix by explicitly utilizing the label information of the training samples. Take a collection of n training samples $X=[x_1,x_2,\cdots,x_n]\in\mathbb{R}^{m\times n}$ from c classes that have a corresponding zero-one label matrix of $H=[h_1,h_2,\cdots,h_n]\in\mathbb{R}^{c\times n}$, where h_i is the label vector for x_i (i.e., if x_i is from the k-th $(k=1,2,\cdots,c)$ class, then only the k-th entry of h_i is 1, while the others are all 0). During the training phase of supervised feature extraction, the samples features are usually fed into a linear classifier $f(x,W)=W^Tx$. By doing so, the model parameters W can be obtained. Accordingly, we define a supervised learning function that smoothly integrates the label information matrix H and linear classifier f as follows:

$$\min_{Q,W}\sum_{i=1}^n\varphi(h_i,f(Q^Tx_i,W))+\lambda_4\|W\|_F^2\tag{13}$$

where φ is the classification loss function, $\lambda_4>0$ is a regularization parameter, and $f(Q^Tx_i,W)$ indicates that the embedding subspace samples are fed into the linear classifier f. In our paper, we apply the parameter of the linear classifier f, which is Q, to learn the discriminative projection (i.e., $W=Q$). In this way, we seek to optimize Q by minimizing the loss of the classification error so that the extracted discriminative features are tightly coupled with the classification purpose. Hence, $\Psi(X,Q)$ of problem (12) is defined as follows:

$$\Psi(X,Q)=\min_Q\sum_{i=1}^n\varphi(h_i,f(Q^Tx_i,Q))+\lambda_4\|Q\|_F^2.\tag{14}$$

Similar to (11), to encourage the projection matrix Q in $\Psi(X,Q)$ to select the most discriminative features, the Frobenious norm is replaced by the $\ell_{2,1}$ norm, and we set φ as the multivariate rigid regression. Therefore, we have

$$\Psi(X,Q)=\frac{1}{2}\|H-Q^TX\|_F^2+\lambda_4\|Q\|_{2,1}.\tag{15}$$

By incorporating (15) into (12), the final objective function for robust feature extraction is defined as follows:

$$\begin{aligned}&\min_{Z,E,R,Q}\lambda_1\|Q\|_{2,1}+\lambda_2\|E\|_{2,1}+\lambda_3\|Z\|_*+\frac{1}{2}\|H-Q^TX\|_F^2\\&s.t.\ X=RQ^TXZ+E, R^TR=I.\end{aligned}\tag{16}$$

Model (16) is named latent low-rank and sparse embedding (LLRSE).

B. Optimization

In this section, we solve problem (16) by using the alternating direction method of multipliers (ADMM) [36]. A convergence analysis of ADMM for nonconvex problems is provided in [37].

We first introduce an auxiliary variable A and a constraint $Z=A$ to make (16) separable. As a result, we can relax (16) as follow:

$$\begin{aligned}&\min_{Z,E,R,Q,A}\lambda_1\|Q\|_{2,1}+\lambda_2\|E\|_{2,1}\\&\qquad+\lambda_3\|A\|_*+\frac{1}{2}\|H-Q^TX\|_F^2\\&\qquad s.t.\ X=RQ^TXZ+E, R^TR=I, Z=A.\end{aligned}\tag{17}$$

Furthermore, the augmented Lagrangian function of (17) is defined by

$$\begin{aligned}\mathcal{L}(Z,E,R,Q,A)&=\lambda_1\|Q\|_{2,1}+\lambda_2\|E\|_{2,1}\\&+\lambda_3\|A\|_*+\frac{1}{2}\|H-Q^TX\|_F^2\\&+<Y_1,X-RQ^TXZ-E>+<Y_2,Z-A>\\&+\frac{\beta}{2}(\|X-RQ^TXZ-E\|_F^2+\|Z-A\|_F^2)\\&s.t.\ R^TR=I\end{aligned}\tag{18}$$

where $< A, B >= trace(A^T B)$, β is a penalty parameter, and Y_1 and Y_2 are two Lagrange multipliers.

By using some mathematics tricks, (18) can be reformulated as follows:

$$\begin{aligned}\mathcal{L}(Z,E,R,Q,A) &= \lambda_1\|Q\|_{2,1} + \lambda_2\|E\|_{2,1} + \lambda_3\|A\|_* \\ &+ \frac{1}{2}\|H - Q^T X\|_F^2 - \frac{1}{2\beta}\|Y_1\|_F^2 - \frac{1}{2\beta}\|Y_2\|_F^2 \\ &+ \frac{\beta}{2}(\|X - RQ^T XZ - E + \frac{Y_1}{\beta}\|_F^2 + \|Z - A + \frac{Y_2}{\beta}\|_F^2) \\ &s.t.\ R^T R = I.\end{aligned} \tag{19}$$

To solve (19), we alternately update these variables Z, W, R, Q, and A.

Step 1. Update Q: fix Z, E, R, A, the problem w.r.t. Q becomes

$$\begin{aligned}\min_Q \lambda_1\|Q\|_{2,1} &+ \frac{1}{2}\|H - Q^T X\|_F^2 \\ &+ \frac{\beta}{2}(\|X - RQ^T XZ - E + \frac{Y_1}{\beta}\|_F^2).\end{aligned} \tag{20}$$

Taking the derivative of $\mathcal{L}(Q)$ w.r.t. Q and setting it to zero, i.e.,

$$\begin{aligned}\frac{\partial \mathcal{L}(Q)}{\partial Q} =& 2\lambda_1 DQ + \beta(XZZ^T X^T Q - XZG^T R) \\ &+ (XX^T Q - XH^T) = 0\end{aligned} \tag{21}$$

where $G = X - E + \frac{Y_1}{\beta}$, and D is a diagonal matrix with diagonal elements $D_{ii} = \frac{1}{2\|q^i\|_2}$ $(i = 1, 2 \cdots, m)$, i.e.,

$$\begin{bmatrix} \frac{1}{2\|q^1\|_2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{2\|q^m\|_2} \end{bmatrix}$$

q^i is the i-th row of Q. When $\|q^i\|_2 = 0$, we let $D_{ii} = 0$, otherwise, $D_{ii} = \frac{1}{2\|q^i\|_2}$. Therefore, we obtain

$$Q^* = (\lambda_1 D + XX^T + \beta XZZ^T X^T)^{-1}(XH^T + \beta XZG^T R). \tag{22}$$

Step 2. Update A: fix Z, E, R, Q and update A by minimizing the following subproblem:

$$\min_A \lambda_3\|A\|_* + \frac{\beta}{2}\|Z - A + \frac{Y_2}{\beta}\|_F^2. \tag{23}$$

Problem (23) has a closed form solution by using the Singular Value Thresholding (SVT) operator [38]. Specifically, given a arbitrary matrix $A \in \mathbb{R}^{m\times n}$ with rank r, the singular value decomposition (SVD) of it is

$$A = U_{m\times r}\Lambda V_{n\times r}^T, \Lambda = diag(\sigma_1, \sigma_2, \cdots, \sigma_r) \tag{24}$$

where $\sigma_1, \sigma_2, \cdots, \sigma_r$ are positive singular values, and $U_{m\times r}$ and $V_{n\times r}$ are corresponding left-singular vectors and right-singular vectors of A, respectively. For a given $\tau > 0$, the singular value shrinkage operator is defined as follows:

$$S_\tau(A) = U_{m\times r} diag(\{max(0, \sigma_j - \tau)\}_{1\leqslant j\leqslant r}) V_{n\times r}^T \tag{25}$$

Then, we get

$$A^* = S_{\frac{\lambda_3}{\beta}}(Z + \frac{Y_2}{\beta}) \tag{26}$$

which can be solved by SVT algorithm[1].

Step 3. Update R: fix Z, E, Q, A, and drop the irrelevant terms with respect to R, the subproblem for updating R is

$$\begin{aligned}&\min_{R:R^T R=I} \frac{\beta}{2}\|X - RQ^T XZ - E + \frac{Y_1}{\beta}\|_F^2 \\ &= \min_{R:R^T R=I} \frac{\beta}{2}\|G^T - Z^T X^T QR^T\|_F^2\end{aligned} \tag{27}$$

where $G = X - E + \frac{Y_1}{\beta}$. The problem (27) is a orthogonal procrustes problem (OPP) [39] problem, whose solution can be directly obtained from SVD, i.e.,

$$U_{m\times m}\Lambda V_{d\times d}^T = GZ^T X^T Q \tag{28}$$

Then, we get

$$R^* = UV^T \tag{29}$$

where U and V are the left-singular vectors and right- singular vectors of $GX^T Q$, respectively.

Step 4. Update E: fix Z, A, R, Q, the subproblem for updating E is

$$\min_E \lambda_2\|E\|_{2,1} + \frac{\beta}{2}\|X - RQ^T XZ - E + \frac{Y_1}{\beta}\|_F^2. \tag{30}$$

Let $G = X - RQ^T XZ + \frac{Y_1}{\beta}$, (30) is equivalent to

$$\min_E \frac{\lambda_2}{\beta}\|E\|_{2,1} + \frac{1}{2}\|E - G\|_F^2 \tag{31}$$

which can be decoupled as:

$$\min_{e_i} \frac{\lambda_2}{\beta}\|e_i\|_2 + \frac{1}{2}\|e_i - g_i\|_2^2 \tag{32}$$

where e_i and g_i are the i-th column of matrix E and G respectively. The solution to problem (32) is given by

$$e_i = \begin{cases}(1 - \frac{\lambda_2/\beta}{\|g_i\|_2})g_i & \|g_i\|_2 > \lambda_2/\beta \\ 0 & \|g_i\|_2 \leq \lambda_2/\beta.\end{cases} \tag{33}$$

Step 5. Update Z: fix A, R, Q, E, the subproblem for updating Z is

$$\min_Z \frac{\beta}{2}\|X - RQ^T XZ - E + \frac{Y_1}{\beta}\|_F^2 + \frac{\beta}{2}\|Z - A + \frac{Y_2}{\beta}\|_F^2. \tag{34}$$

Defining $G_1 = X - E + \frac{Y_1}{\beta}$ and $G_2 = A - \frac{Y_2}{\beta}$, and setting the derivative of $\mathcal{L}(Z)$ w.r.t Z to zero, we have

$$\begin{aligned}\mathcal{L}(Z) &= tr[(Z^T X^T QR^T - G_1^T)(RQ^T XZ - G_1)] \\ &+ tr[(Z^T - G_2^T)(Z - G_2)] \\ &= tr(Z^T(X^T QQ^T X + I)Z - 2Z^T(X^T QR^T G_1 + G_2))\end{aligned} \tag{35}$$

Let $\frac{\partial \mathcal{L}(Z)}{\partial Z} = 0$, we have

$$(X^T QQ^T X + I)Z - (X^T QR^T G_1 + G_2) = 0. \tag{36}$$

From (36), it is obvious that

$$Z^* = (X^T QQ^T X + I)^{-1}(X^T QR^T G_1 + G_2). \tag{37}$$

[1]http://svt.stanford.edu/index.html

Step 6. Update Y_1, Y_2 and β: Y_1, Y_2, and β are respectively updated by

$$\begin{aligned} Y_1 =& Y_1 + \beta(X - RQ^T XZ - E) \\ Y_2 =& Y_2 + \beta(Z - A) \\ \beta =& \min(\rho\beta, \beta_{max}). \end{aligned} \tag{38}$$

The detailed optimization procedure of LLRSE is presented in Alg. 1.

Algorithm 1 Solving problem (16) via ADMM

Input: Data matrix X, the max number of iteration $maxIter$, and three penalty parameters λ_1, λ_2 and λ_3.

Output: Discriminative projection matrix Q, orthogonal matrix R for reconstruction, low-rank matrix Z, and error matrix E.

Initialize: Set $t = 0$; Initialize $Q^0 = \mathbf{0}$, $A^0 = \mathbf{0}$, $R^0 = \mathbf{0}$, $E^0 = \mathbf{0}$, $Z^0 = \mathbf{0}$, $Y_1^0 = \mathbf{0}$, $Y_2^0 = \mathbf{0}$, and set $\beta = 10^{-1}$, $\beta_{max} = 10^5$, $\rho = 1.1$, $\varepsilon_1 = 10^{-6}$, $\varepsilon_2 = 10^{-6}$.

Compute the label matrix $H \in \mathbb{R}^{c\times n}$ of X;
repeat
  Given A^t, R^T, E^t, Z^t, and update the projection matrix Q^{t+1} according to (22);
  Given Q^{t+1}, R^T, E^t, Z^t, and update the the auxiliary matrix A^{t+1} according to (26);
  Given $A^{t+1}, Q^{t+1}, E,^t Z^t$, and update the reconstruction matrix R^{t+1} according to (29);
  Given $A^{t+1}, R^{t+1}, Z^t, Q^{t+1}$, and update the error matrix E^{t+1} according to (33);
  Given $A^{t+1}, R^{t+1}, E^{t+1}, Q^{t+1}$, and update low rank matrix Z^{t+1} according to (37);
  Update the Lagrange multipliers Y_1^{t+1} and Y_2^{t+1} and penalty variable β according to (38);
  Check the convergence conditions
  $\|X - R^{t+1}(Q^{t+1})^T XZ^{t+1} - E^{t+1}\|_\infty \le \varepsilon_1$,
  $\|Z^{t+1} - A^{t+1}\|_\infty \le \varepsilon_1$,
  $(\mathcal{L}^{t+1} - \mathcal{L}^t) \le \varepsilon_2$;
  Update t = t+1.
until Convergence criterion is satisfied, or $t > maxIter$.

C. Complexity Analyses

The computational complexity of LLRSE mainly depends on the complexity of Alg. 1 in which the main time-consuming operations are solving the inverse of square matrix and performing SVD. To be precise, the computational complexity of the inverse operation of an $m \times m$ square matrix is $O(m^3)$, and the computational complexity of the SVD operation of an $m \times n$ matrix is $O(n^3)$.

Next, we discuss the main computational complexity of each step. To update Q, the most demanding computation is the matrix inverse transformation. Given a data matrix $X \in \mathbb{R}^{m\times n}$, the learned projection matrix $Q \in \mathbb{R}^{m\times d}$ corresponding to the diagonal matrix $D \in \mathbb{R}^{m\times m}$ is involved in computing the inverse operation of $\lambda_1 D + XX^T + \beta XZZ^T X^T$. Thus, the computational complexity is about $O(m^3)$. To update A, the most demanding computation is the SVD computation of matrix $A \in \mathbb{R}^{n\times n}$, and thus the computational complexity is $O(n^3)$. To update R, the computational complexity of the SVD computation of $(X - E + \frac{Y_1}{\beta})X^T Q \in \mathbb{R}^{m\times d}$ is $O(d^3)$. To update E, we only need to visit matrix $E \in \mathbb{R}^{m\times n}$ row by row; thus, the computational complexity is $O(m)$. To update Z, the most demanding computation is the inverse transformation of matrix $X^T QQ^T X \in \mathbb{R}^{n\times n}$, and thus the computational complexity is $O(n^3)$. Combined, the computational complexity of each iteration of LLRSE is about to $O(m^3 + n^3 + d^3 + m + n^3)$. In practice, we suppose that $m \leqslant n$ and the value of d is small. Hence, the total computational complexity of LLRSE can be loosely thought of as $O(tn^3)$, where t is the number of iterations.

IV. Experimental Results and Analysis

We evaluate the accuracy, robustness, parameter sensitivity, and convergence of the proposed LLRSE using six widely used datasets, including ORL [4, 12], Extended Yale B [30], AR [12, 19], YouTube-Celebrities [40], USPS [30], and COIL-20 [41]. Among them, the first four are face datasets that contain face expressions, lighting conditions, decorations, random pixel corruptions, and block occlusions. They are used to test face recognition performance of LLRSE. The USPS and COIL-20 datasets are used to test the performance of LLRSE in handwritten digital recognition and object recognition, respectively.

Our experimental platform is MATLAB R2016b operating on a MacBook Pro 2017 with an Intel Core i5 (2.3GHz) CPU and 8G RAM. If the readers are interested in the source code of our proposed method, please contact the first author by email.

A. Datasets

Four face datasets: The four face datasets are widely employed in evaluating the algorithmic performance [12, 19, 30, 40].

The Cambridge ORL[2] face dataset contains 400 face images of 40 individuals under dark ground. All the images are resized[3] to 32×32 pixels. Some samples of 8 subjects are shown in Fig. 1a.

The Extended Yale B (EYaleB)[4] dataset contains 2,414 frontal face images of 38 individuals with small variations in facial expressions and head poses. Each individual has about 64 images taken under various controlled illumination conditions. The face region of each image is cropped and resized to 32×32 pixels. Some samples of the first 8 individuals are shown in Fig. 1b.

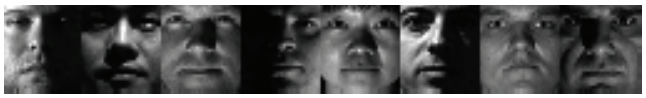
The AR[5] dataset contains over 4,000 frontal face images of 126 individuals (56 female and 70 male). Each individual is involved in 2 sessions, and each individual has 13 images taken in each session. Among the 13 images, there are 3 images with sunglasses, another 3 images with scarves, and the remaining 7 are neutral images with different exaggerated facial expressions, myopia glasses, and illuminations under

[2]https://www.cl.cam.ac.uk/research/dtg/attarchive/facedataset.html
[3]The nearest interpolation strategy is used.
[4]http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html
[5]http://www2.ece.ohio-state.edu/ aleix/ARdataset.html

different angle. We crop the face region of each image and resize it to 60×44 pixels. The face images of 8 different individuals are shown in Fig. 1c.

The YouTube-Celebrities (YTC)[6] dataset contains 1,910 video clips of 47 celebrities from YouTube. These videos are obtained from real-life scenarios, where there are very large variations in face expressions, appearances and poses, as well as noise, misalignments, poor quality, and occlusions. Each of the 1910 video clips contains 8-400 frames, and each celebrity has more than 1,000 images. The face images are resized to 30×30 pixels. Some face images of different individuals are shown in Fig. 1d. For our experiments, 100 images of each celebrity are randomly selected as the experimental data (i.e., 4,700 images in total).

(a) ORL



(b) EYaleB

(c) AR

(d) YTC

Fig. 1: Samples of the first 8 individuals from the four face datasets.

A digit dataset: The USPS[7] [30] handwriting digital image dataset contains 10 categories from "0" to "9". Each category contains 1,100 examples. We resize each image to be the size of 16×16 pixels. Fig. 2 shows some samples.

Fig. 2: Samples from the USPS handwritten digital dataset.

An object dataset: The COIL-20[8] [41] dataset contains 1,440 images of 20 objects (such as a cat and car model), each of which has 72 images with different views taken at pose intervals of 5° against a black background. For our experiments, each image is converted to 32×32 pixels. Fig. 3 shows some kawaii images from 8 different classes.

[6]http://www.cs.tau.ac.il/ wolf/ytfaces/

[7]http://www.cad.zju.edu.cn/home/dengcai/Data/data.html

[8]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

Fig. 3: Some kawaii samples from the COIL-20 object dataset.

It should be note that although the subject disjoint protocol[9] [42] can avoid positive bias in recognition results better, the sampling strategy of this paper does not follow it since this paper concerns with feature extraction rather than traditional recognition tasks. Analogously to [12, 19, 20, 28–30], all images of any subject are divided into two groups, one for extracting features and the other for evaluating the abilities of extracted features. In addition, as far as we know, it is very necessary to normalize each sample to the unit norm to remove amplitude variations and only focus on the underlying distribution shape. In this paper, ℓ_2 normalization (least square error) is employed to normalize the data matrix X.

B. Compared Methods

We compare LLRSE with some representation-based methods, including SRC [12], CRC [28], LRC [11], LRRC [19], LatLRR [20], SALPL [29], and LRE [30]. The codes of all compared methods that have been released by the original authors are used, and the suggested parameters are adopted from their published papers.

For the three baseline methods, SRC, CRC and LRC, all the training samples themselves are employed as the reconstruction dictionary. The minimal representation residual $\min_i r_i(y) = \|y - X\delta_i(\hat{\alpha})\|_2$ is used for classification, where y is a query image, and $\delta_i(\cdot)$ is a vector whose only nonzero entries are the entries in $\hat{\alpha}$ that are associated with class i. For LRRC, the learned representation matrix Z is fed into a linear regression classifier to learn a linear classifier $\hat{W}$, i.e.,

$$\hat{W} = \arg\min_W \|H - WZ\|_2^2 + \lambda\|W\|_2^2 \tag{39}$$

where $H \in \mathbb{R}^{c\times n}$ is the zero-one label matrix of X. The problem (39) has a closed form solution as

$$\hat{W} = HZ^T(ZZ^T + \lambda I)^{-1}. \tag{40}$$

After obtaining the linear classifier $\hat{W}$, we then compute the representation matrix Z of the testing data $Y = [y_1, \cdots, y_q] \in \mathbb{R}^{m\times q}$ so the class label of a query image y_i is $l = \arg\max_l(\hat{W}z_i)$. For the unsupervised LRE, similar to [30], we extend LRE to the supervised scenario where the label information is available for training samples. By integrating the label information into LRE, the supervised LRE model is defined as follows:

$$\arg\min_{Z,Q} \|Z\|_* + \lambda\|Q^TX - Q^TXZ\|_{2,1} + \Psi(X,Q) \tag{41}$$

where $\Psi(X,Q) = \frac{1}{2}\|H - Q^TX\|_F^2$, and H is also the zero-one label matrix of X. Once the projection matrix Q is obtained by our LLRSE, LRE, SALPL, or LatLRR, the original samples

[9]Subject disjoint protocol means that all images from a single subject are either in the training set or test set.

are projected into the low-dimensional subspace and the classification result is directly obtained via a 1-Nearest Neighbor (1-NN) classifier. (Note that although the L of the LatLRR does not reduce the dimensionality, it can also be regarded as a projection matrix.)

Before presenting the experiments and results, we first present the detailed implementation of the proposed LLRSE. The reconstruction matrix R and projection matrix Q are initialized by performing PCA on the data matrix and zeros matrix, respectively. The parameters are set to $\lambda_1 = 0.1$, $\lambda_2 = 0.001$, $\lambda_3 = 0.001$, $\rho = 1.01$, $\beta = 0.1$, $\beta_{max} = 10^5$, $\varepsilon_1 = 10^{-6}$, $\varepsilon_2 = 10^{-6}$, and $maxIter = 30$. In particular, since there is a classification loss term $\|H - Q^T X\|_F^2$ in (16), we need to set the reduced dimensionality of Q to be the same as the number of the training classes, i.e., c. Moreover, for fair comparison in each round, every experiment of the compared methods runs 10 times (unless otherwise stated), and then the average recognition accuracy with the standard deviation of all runs is reported.

C. Recognition Accuracy

To evaluate the performance of the learned features from LLRSE, we perform some face, object, and handwritten digit recognition tasks in turn.

1) Face recognition evaluation: We evaluate LLRSE on four face datasets, EYaleB, AR, ORL, and YTC. For EYaleB, each individual has 64 face images. We randomly select $N = \{10, 15, 20, 25\}$ images for training, and the remainders are the testing samples. For AR, there are 7 neutral face images without occlusions for each individual in each session. This results in 14 neutral images per individual. We randomly select $N = \{3, 4, 5\}$ neutral images of each individual for training, and the remaining neutral images are used for testing. For ORL, there are 10 images of each individual. We randomly select $N = \{5, 6, 7\}$ images from each person for training, and the remaining images are used for testing. For YTC, there are 100 face images for each celebrity. We randomly select $N = \{10, 20, 30, 40\}$ images from each celebrity for training, and the remaining images are used for testing. After running each method 10 times, the average recognition accuracies and standard deviations of all the compared methods on these datasets are calculated and summarized in Table I. The experimental results indicate that LLRSE achieves the best performance on all four datasets with the varying number of training samples. For instance, compared to the relevant LRE method, average improvements of 1.88%, 3.27%, 2.92%, and 2.66% improvements are achieved by LLRSE on AR, ORL, EYaleB, and YTC datasets, respectively. Similarly, LLRSE is visibly better than SALPL, LatLRR, and LRRC.

Furthermore, the receiver operating characteristics (ROC) curve is useful for evaluating classifiers and visualizing classification performance. Therefore, we employ the ROC curve to evaluate the learned features. Due to spatial limitations, we only plot the ROC curve of the compared methods for the YTC dataset since it is a real-life dataset that is very representative. The results are shown in Fig. 4. It can be seen that the proposed method always achieves a high true positive rate (TPR) at a low false positive rate (FPR), and it also achieves the highest area under the ROC curve (AUC) values.

2) Object recognition evaluation: We utilize the COIL-20 dataset to evaluate the performance of LLRSE for object recognition task. For each class of COIL-20, 10 samples are randomly selected from each category for training and the remaining images are used for testing. The average accuracies are reported in Table II. It is observed that LLRSE achieves the highest recognition accuracy compare to all the other methods. The results demonstrate that LLRSE is also applicable to extracting object images's features. In fact, only three categories obtain relatively bad recognition rates: object 2 is 73%, object 5 is 69%, and object 6 is 71%. Further, the other categories are all classified well, some of which (object 8, object 10 to object 18, and object 20) can reach as high as 100% recognition accuracy.

3) Handwritten Digit Recognition Evaluation: We also test LLRSE for the recognition of handwritten digits by using the USPS dataset. We randomly choose 40 samples per category for training, and the remaining images are used for testing. The experimental results are summarized in Table III. We see that LLRSE again outperforms the other methods. In fact, the best recognition accuracy (95%) is achieved for recognizing '0', and the poorest accuracy (80%) is achieved for recognizing '8' since it can easily be mistaken for '3'.

D. Robustness and Recovery Capability

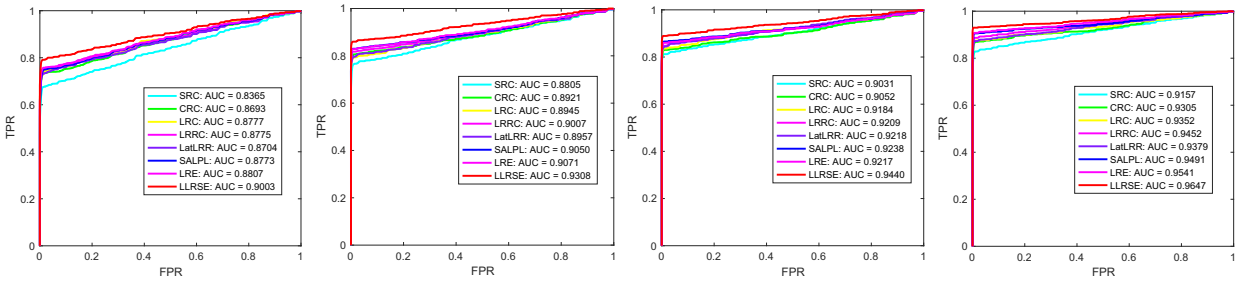
To evaluate the robustness and recoverability of LLRSE against multiple types of corruption and noise, we further perform some additional experiments and present the results for the EYaleB and AR datasets.

1) The AR dataset with occlusion: In real-life applications, the vast majority of real data are corrupted in various ways. Hence, we evaluate LLRSE under complex scenarios, such as those in real-life environments. The AR dataset is undoubtedly the best choice since the unobscured images contain varying illuminations and expressions, and obscured face images are corrupted by occlusion from sunglass (approximatively 20% of the area of the face image) or scarf (approximatively 40% of the area of the face image). Fig. 5(a) shows some corrupted images from AR. In the following experiments, we design three different complex scenarios based on the AR dataset, including (1) obscured by sunglass scenario, (2) obscured by scarf scenario, and (3) a mixed (sunglass & scarf) scenario. The experimental training samples are set as in Table IV, and the remaining samples are used for testing. The detailed comparison results are reported in Fig. 6. It can be seen that LLRSE achieves the best classification accuracies in all types of corruption, with average accuracies that are 2.82% and 3.34% higher than the second best method SALPL and the relevant LRE method, respectively.

2) The EYaleB dataset with random pixel corruption: To evaluate the robustness of LLRSE to different levels of random pixel corruption, we randomly select 20 images from each individual for training and the remaining images are used for testing. For the 5 of the 20 images, we randomly choose pixels from each image and corrupt them using salt and pepper noise.

TABLE I: Average recognition accuracies (%) with standard deviations (%) of the compared methods on four face datasets.

datasets	N	Methods							
		SRC	CRC	LRC	LRRC	LatLRR	SALPL	LRE	**LLRSE**
AR	3	77.40±0.99	76.92±1.14	77.55±1.45	78.31±0.76	76.81±0.77	78.90±0.82	78.84±0.75	**80.22±0.64**
	4	84.71±0.76	84.52±0.87	85.00±1.20	85.96±0.70	84.13±0.98	85.12±0.84	85.29±0.62	**87.31±0.62**
	5	92.86±0.76	91.88±0.86	93.45±1.04	93.77±0.44	91.25±0.71	94.09±0.70	93.18±0.60	**95.43±0.41**
ORL	5	88.21±0.66	88.79±0.64	84.75±0.63	85.19±0.72	86.20±0.44	91.71±0.52	89.13±0.42	**92.68±0.28**
	6	91.22±0.44	90.85±0.38	89.92±0.62	90.33±0.61	90.05±0.32	93.77±0.42	92.20±0.28	**95.20±0.25**
	7	94.32±0.38	94.10±0.29	94.56±0.39	92.50±0.57	92.58±0.21	95.41±0.35	95.07±0.20	**98.33±0.12**
EYaleB	10	87.91±0.49	83.45±1.46	87.28±1.18	83.51±1.23	82.25±1.31	86.24±0.94	84.71±0.74	**89.21±0.34**
	15	93.64±0.52	90.21±1.28	91.10±0.94	90.33±0.83	89.17±0.94	93.72±0.88	92.86±0.60	**95.31±0.25**
	20	95.28±0.45	93.44±0.89	92.36±0.68	91.95±0.91	90.27±0.62	96.06±0.69	94.36±0.62	**97.31±0.30**
	25	96.32±0.36	96.13±0.75	96.21±0.64	96.18±0.86	93.57±0.66	97.15±0.72	96.75±0.51	**98.56±0.24**
YTC	10	67.24±1.90	72.95±2.21	74.23±2.10	74.86±1.98	72.89±1.76	74.56±1.68	76.21±1.56	**78.98±1.40**
	20	76.47±1.72	80.16±1.91	79.10±1.82	81.36±1.88	79.47±1.58	82.62±1.44	82.58±1.53	**85.66±1.22**
	30	80.78±1.35	82.63±1.41	83.45±1.55	84.62±1.62	84.29±1.60	86.22±1.44	85.79±1.28	**88.76±1.06**
	40	82.76±1.29	86.79±1.42	87.48±1.34	88.58±1.38	87.18±1.33	90.48±1.39	90.71±1.11	**92.93±0.89**



(a) $N = 10$. (b) $N = 20$. (c) $N = 30$. (d) $N = 40$.

Fig. 4: Mean ROC curves of the compared methods under different N on the YTC dataset.

TABLE II: Recognition accuracies (%) of the compared methods on the COIL-20 dataset.

SRC	CRC	LRC	LRRC	LatLRR	SALPL	LRE	**LLRSE**
89.35	84.37	88.90	80.79	88.97	91.37	91.93	**93.79**

TABLE III: Recognition accuracies (%) of the compared methods on the USPS dataset.

SRC	CRC	LRC	LRRC	LatLRR	SALPL	LRE	**LLRSE**
82.38	84.42	79.14	76.75	87.26	88.98	88.46	**90.79**

(a) The AR dataset with sunglasses and scarves occlusions.

(b) The EYaleB dataset with random pixel corruptions and block occlusions, the first five images represent samples with corrupted pixels under different percentages (i.e., 0%, 10%, 20%, 30%, and 40%), the last three images represent examples with block occlusions under different percentages (i.e., 5%, 15%, and 20%).

Fig. 5: Some corrupted samples from the AR and EYaleB datasets.

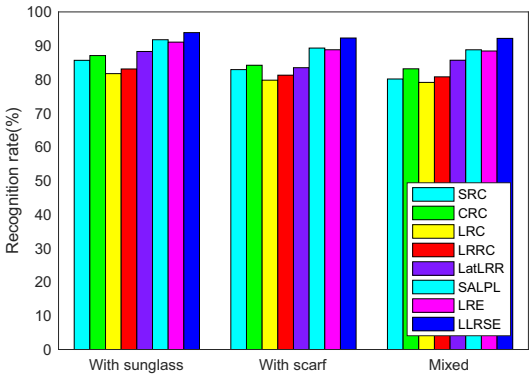


Fig. 6: Recognition accuracies (%) of the compared methods on the AR dataset with three different types of corruptions.

The percentage of corrupted pixels varies from 10% to 40%. The first five images of Fig. 5(b) show some corrupted images with different percentages of pixel corruption from the EYaleB dataset. The results are reported in Table V. It can be seen that LLRSE again dramatically outperforms others. Moreover, it is obvious that the trend of the accuracy degradation is not as obvious as the increasing of corruption ratio; hence, its advantages become more obvious.

3) The EYaleB dataset with block occlusion corruption: To evaluate the robustness of LLRSE to different levels of block occlusion corruption, we randomly selected 20 images from

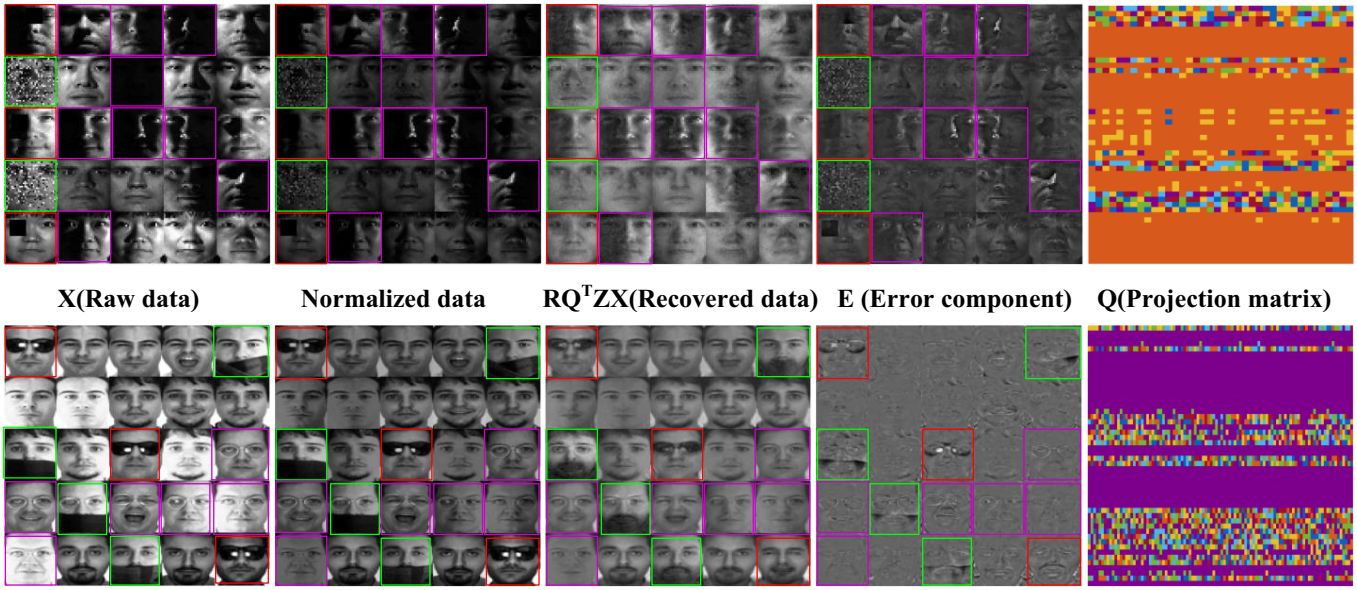


Fig. 7: Recovery results of LLRSE on the EYaleB (the first row) and AR (the second row) datasets. In terms of columns, from left to right, are original training data, normalized data, recovered data, error data, and projection matrix in turn. Specifically, for EYaleB, the green rectangle represents images with random pixel corruptions, the red rectangle represents the images with block occlusion corruptions, and the purple rectangle represents the images with obvious illumination and facial expression variations. For AR, the green rectangle represents the face images with scarves, the red rectangle represents the face images with sunglasses, and the purple rectangle represents the images with myopic glasses. And we only show the first 50 rows of their respective projection matrices Q for comparison. It can be seen that some rows of the learned projection matrix Q are zeros due to the row-sparsity property of $\|\cdot\|_{2,1}$-norm. Therefore, $\|\cdot\|_{2,1}$-norm has better interpretability for the importance of features.

TABLE IV: Experimental training samples settings for each individual on the AR dataset. (Note: S1=Session 1, S2=Session 2, G=Sunglass, S=Scarf, and N=Neutral. For example, S1∩G represents the samples with sunglass in session 1).

Scenarios	S1∩G	S1∩S	S1∩N	S2∩G	S2∩S	S2∩N
Sunglasses	2	0	4	2	0	4
Scarves	0	2	4	0	2	4
Mixed	1	1	4	1	1	4

TABLE V: Recognition accuracies (%) of the compared methods on the EYaleB dataset with different percentages of random pixel corruption.

Ratio(%)	SRC	CRC	LRC	LRRC	LatLRR	SALPL	LRE	**LLRSE**
10	92.36	92.49	92.04	90.83	89.16	93.37	94.28	**95.72**
20	88.96	89.00	89.59	87.98	84.32	91.82	90.79	**94.28**
30	84.32	82.21	86.93	80.37	76.82	88.71	88.21	**92.41**
40	80.12	81.43	81.86	78.76	67.85	83.14	83.56	**91.86**

TABLE VI: Recognition accuracies (%) of the compared methods on the EYaleB dataset with different percentages of random block occlusion.

Ratio(%)	SRC	CRC	LRC	LRRC	LatLRR	SALPL	LRE	**LLRSE**
5	92.53	92.18	91.24	90.13	89.16	94.21	94.02	**96.12**
10	90.02	88.64	88.55	86.40	85.62	92.23	93.18	**95.44**
20	86.79	86.61	85.29	85.24	78.84	90.89	91.44	**93.91**
30	80.18	80.44	84.46	74.68	70.21	88.79	88.62	**92.62**

each individual of the EYaleB dataset for training, and the remaining images are used for testing. Among these images, the 5 of the 20 images are corrupted by block occlusions, which is randomly added to different locations in each image with different percentages that vary from 5% to 30%, respectively. Fig. 5(b) shows some corrupted examples (the last three images). The results are tabulated in Table VI, which shows that our LLRSE is robust to block occlusion corruptions. More precisely, the recognition accuracies of LLRSE are the best for the whole period: they are approximatively 1.91%, 2.05%, 1.52%, and 3.07% higher than the second best method, SALPL, when the block occlusion percentages are 5%, 10%, 20%, and 30%, respectively. Compared to the most relevant LRE method, the improvements of LLRSE are 2.1%, 2.26%, 2.47%, and 4.00% when the block occlusion percentages are 5%, 10%, 20%, and 30%, respectively.

4) Recoverability against corruption with the EYaleB and AR datasets: The main goal of the proposed method is to reduce the dimensionality such that the reduced discriminative features can represent and reconstruct the clean original data as much as possible. Thus, we illustrate the effect of recovering clear images from the corrupted images with different types of corruption. We randomly select 20 images per person for training, and the sample production strategy follows the rules that are given as follows. For EYaleB, each person has 5 images with 40% random pixel corruption and 5 images with 10% block occlusion corruption. For AR, each person has 4 images with sunglasses, and 4 images with scarves. In fact, a large

percentage of these images inherently contain different facial expressions, illumination conditions, and myopia glasses. To visually view the reconstruction results, we visualize the data vectors as 32×32 and 44×60 gray images for the EYaleB and AR datasets, respectively. The original images X, the normalized images, the recovered images RQ^TXZ, the error component E, and the projection matrix Q are visually shown in Fig. 7. To highlight the advantages of the recovered results under different corruption, the visual images of X, RQ^TXZ, and Z contain three different types of colored solid rectangles, where different colors represent different types of corruption.

From Fig. 7, we have the following three observations. (1) From the results on the EYaleB, we see clearly that our LLRSE can effectively recover the facial expressions, illumination shadows, random pixel corruptions, and random block occlusions. (2) From the results on the AR, we see that LLRSE can well recover the facial expressions, slightly occlusions from myopia glasses, and serious occlusions from sunglasses and scarves. (3) From the last column of Fig. 7, for both on the EYaleB and AR datasets, we see clearly that the projection matrix Q has the row-sparsity property, i.e., most of its rows are zero vectors. (4) It can be derived that the learned projection matrix Q has the ability to perform feature selection as well as feature extraction.

E. Experiment Analysis and Discussions

From the effectiveness and robustness experimental results on six widely used datasets, the following observations are achieved.

1) The proposed LLRSE method is greatly superior to other excellent representation-based feature learning methods for face recognition, object recognition, and handwritten digital image recognition. Thus, the learned features are very effective for compressing original high-dimensional data. The main reasons are that (1) LLRSE takes the data reconstruction into account so that the major energy of the original data is retained, (2) the $\|Q\|_{2,1}$ term can effectively remove irrelevant and redundant features, and retain the key ones [43], and (3) a classification loss term is integrated with feature extraction so that the obtained features are discriminative for classification purpose. Hence, LLRSE is competent to perform classification tasks. The related experimental results are shown in Table I, Table II, and Table III.

2) Moreover, for corrupted data, there is no doubt that the robustness of LLRSE is greatly improved. The main reason is that (1) the low-rank constraint can better expose the underlying global structure of samples, which greatly helps us to remove corrupted parts [13]. (2) We adopt the columnwise $\|E\|_{2,1}$ term to model the noise component in data, which encourages the columns of E to be zero since it is assumed that some samples are corrupted (i.e., the corruptions are sample-specific), while the others are clean [44]. Hence, the outlier and noise can be effectively removed. (3) And LLRSE learns a robust and discriminative embedding subspace, which can select and extract the important features, or, in other words, remove the irrelevant and redundant features. In general, among these three reasons, one focuses more on corrupted blocks, another focuses more on outlier samples, and the third focuses more on corrupted features. These advantages apply very well to this case where data usually contain considerable noise and outliers in real-life applications. The related experimental results are shown in Table VI, Table V and Fig. 6.

3) Fig. 7 shows why LLRSE performs so well by visualizing the reconstruction results. The data matrix X is decomposed into RQ^TZX and E. In particular, the error component E represents the noise, and RQ^TZX is used for optimizing the data reconstruction process. Obviously, they can recover the clean part and alleviate the interference of the noise part.

4) In addition to visually evaluating the recoverability of LLRSE in Fig. 7, we give more explanations about the interpretability of features for Q. The component $\|Q\|_{2,1}$ encourages a good row-sparsity property and ranks the importance of the features (i.e., during the training phase, it adaptively assigns large projection weights to the important features, and vice versa). Thus, after obtaining Q, $\|Q^i\|_2$ (Q^i is the i-th row of Q) can be used to quantify the probability of whether the i-th feature is a key one. This demonstrates that the $\ell_{2,1}$-norm is very useful for capturing discriminative features and reducing dimensionality. The experimental results are shown in Fig. 7.

F. Parameter Sensitivity and Convergence

To be identical to the experimental settings in Section IV-C1, we conduct the parameter sensitivity and convergence analyses on the EYaleB ($N = 20$) and AR ($N = 5$) datasets.

1) Parameter sensitivity: To investigate the parameter sensitivity, the variations of the parameters versus the recognition accuracy are tuned. In all experiments, the row-sparsity regularization parameter λ_1 is chosen from the set $QArgs = \{20, 10, 5, 2, 1.5, 0.5, 0.2, 0.1, 0.05, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, the error regularization parameter λ_2 is chosen from the set $EArgs = \{1, 0.1, 10^{-2}, 10^{-3}, 10^{-4}\}$, the low-rank constraint parameter is chosen from the set $LRArgs = \{1, 0.1, 10^{-2}, 10^{-3}, 10^{-4}\}$, and the Lagrange multipliers-related parameters are set to $\rho = 1.01$, $\beta = 0.1$, and $\beta_{max} = 10^5$.

Although we have up to three parameters in our method, we find that λ_1 plays a more important role in the performance, and the other, parameters λ_2 and λ_3, only have a slight effects on the performance for all experiments. In other words, λ_1 is a coarse tuning parameter, and λ_2 and λ_3 are precise tuning parameters. Fig. 9 shows the recognition accuracy (%) versus λ_1 on the EYaleB and AR datasets. When we fix λ_1, then λ_2 and λ_3 are selected from $EArgs$ and $LRArgs$, respectively.

As seen from Fig. 9, some conclusions can be drawn from the results: (1) The best recognition accuracy is not highly sensitive to λ_1 when the value of λ_1 varies from 10^{-1} to 10^{-2}; (2) The recognition accuracy rapidly deteriorates when $\lambda_1 > 0.5$, and slightly deteriorates when $\lambda_1 < 10^{-2}$; and (3) The mean accuracy curves are very close to the respective max accuracy curves, which confirms that our method has high stability and indicates that $\|Q\|_{2,1}$ plays a more important roles in feature learning (i.e., the important features are reserved).

Furthermore, when λ_1 is fixed, we tune λ_2 and λ_3. Fig. 8 shows the recognition accuracy (%) versus the parameters

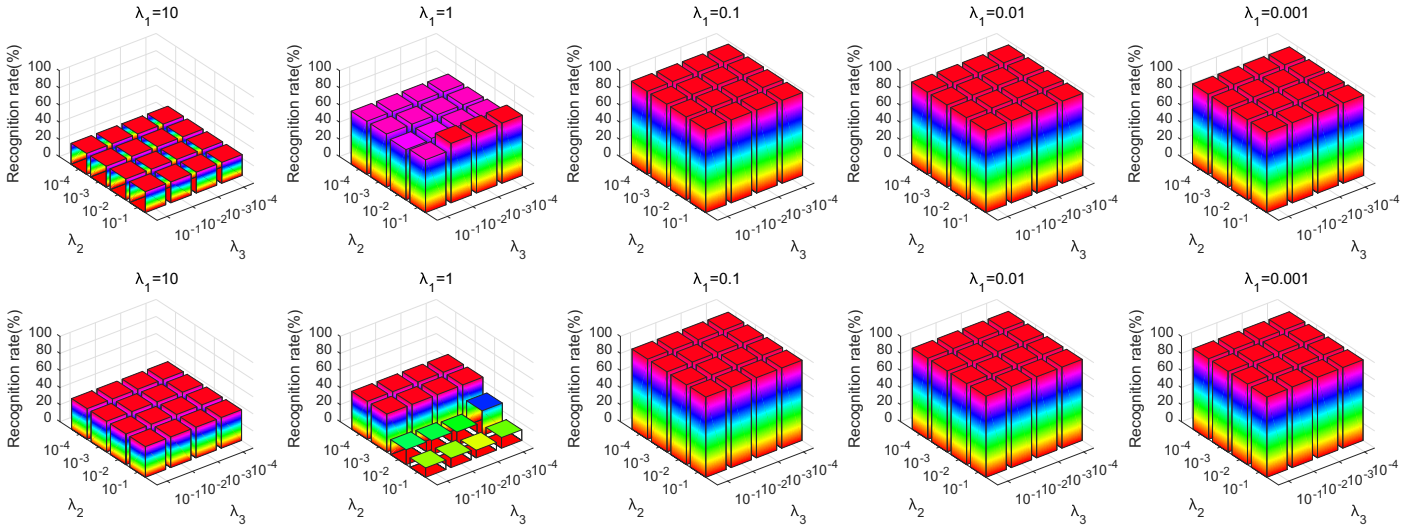


Fig. 8: Recognition accuracies(%) of the proposed method versus the parameters λ_1, λ_2, and λ_3 on the EYaleB (the first row) and AR (the second row) datasets.

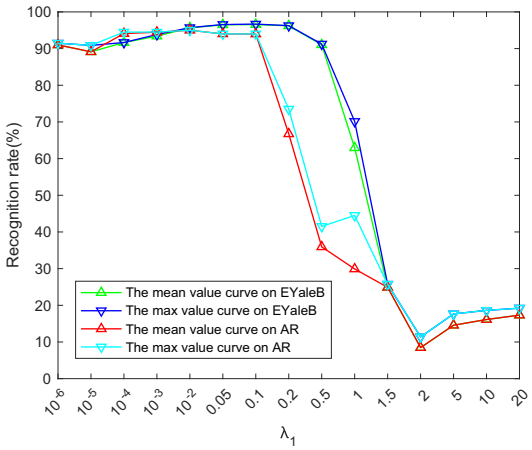


Fig. 9: Recognition accuracy (%) of the proposed method versus λ_1 on the EYaleB and AR datasets. (Note that when λ_1 is chosen, λ_2 and λ_3 are chosen from $EArgs$ and $LRArgs$, respectively.)

λ_2 and λ_3 with a fixed λ_1 on the EYaleB and AR datasets. From all the plots in Fig. 8, it can be seen that (1) the recognition accuracy only slightly varies when λ_1 is fixed; (2) The recognition accuracy is slightly more sensitive to λ_3 than λ_2 when the training images are occluded, such as in the AR dataset. This is because λ_3 controls the low-rank component, and the low-rank assumption is more intuitive than the sparseness assumption for removing structural noise (like the block occlusions); and (3) The recognition accuracy is slightly more sensitive to λ_2 than λ_3 when the training images contain outliers and random noise, such as the EYaleB dataset. This increase in sensitivity occurs because λ_2 controls the error regularization $\|E\|_{2,1}$.

According to the above comprehensive analysis, in this paper, for the sake of simplicity, we fix $\lambda_1 = 0.1$, $\lambda_2 = 0.001$, and $\lambda_3 = 0.001$ for all experiments.

2) Convergence Analysis: Finally, we demonstrate the fast convergence of LLRSE. In Fig. 10, we plot the convergence curves and the difference curves (which are defined as the change of the values between two consecutive iterations) for the EYaleB and AR datasets. The results illustrate that the proposed method converges quickly within approximatively 15 iterations. For space limitation, the convergence of the proposed method for other datasets is not shown in this paper.

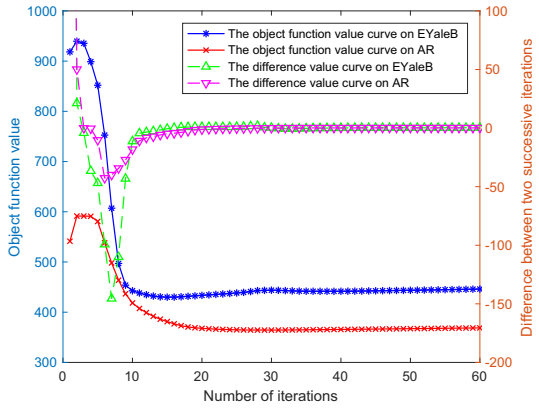


Fig. 10: Convergence curves and difference curves of the proposed method on the EYaleB and AR datasets.

V. Conclusion

In this paper, a robust representation-based method called LLRSE is proposed for image feature extraction, which can address some problems of SR and LRR-based methods, especially for LRE. LLRSE seamlessly integrates an orthogonal data reconstruction term, a classification loss term, and some regularization terms into a unified framework to learn a robust and discriminative embedding subspace. In addition, LLRSE is insensitive to parameters and needs fewer iterations to achieve convergence. Extensive experiments on six widely

used datasets demonstrate that LLRSE is always robust to multiple types of corruption, including illumination variations, block occlusions and pixel corruptions, and it consistently outperforms all the other evaluated state-of-the-art feature learning methods in terms of recognition accuracy and robustness.

In addition to the contents that are presented in this paper, future work consists of the following: 1) Further extension of LLRSE to handle feature learning with large-scale datasets. 2) Application of Schatten p-norm and correntropy to boost the effectiveness and robustness of LLRSE.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (Grant No. 61673220), the Redevelopment Project of Sichuan Defense Science and Technology Office (No. ZYF-2018-106), Sichuan Major Science and Technology Project (No. 2018TZDZX0002), and State Administration of Science project (No. JCKY2017209B010).

The authors would like to thank Jie Wen and Xiaozhao Fang, who shared their RSLDA code on their homepage, and to thank Wai Keung Wong and Zhihui Lai, who shared the LRE code. The authors would also like to thank the anonymous reviewers who provided substantive suggestions for improving our work.

REFERENCES

[1] L. Xie, M. Yin, X. Yin, Y. Liu, and G. Yin, "Low-rank sparse preserving projections for dimensionality reduction," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5261–5274, 2018.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[3] M. Villegas and R. Paredes, "On improving robustness of lda and srda by using tangent vectors," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1094–1100, 2013.

[4] S. Wang and H. Wang, "Unsupervised feature selection via low-rank approximation and structure learning," *Knowledge-Based Systems*, vol. 124, pp. 70–79, 2017.

[5] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1208–1213.

[6] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections (2dlpp) with its application to palmprint recognition," *Pattern recognition*, vol. 40, no. 1, pp. 339–342, 2007.

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

[8] J. Ye and T. Xiong, "Null space versus orthogonal linear discriminant analysis," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 1073–1080.

[9] Y. Zhou and S. Sun, "Manifold partition discriminant analysis," *IEEE transactions on cybernetics*, vol. 47, no. 4, pp. 830–840, 2017.

[10] T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *European conference on computer vision*. Springer, 2008, pp. 725–738.

[11] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[13] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[14] S. Li and Y. Fu, "Learning balanced and unbalanced graphs via low-rank coding," *IEEE Transactions on Knowledge & Data Engineering*, no. 5, pp. 1274–1287, 2015.

[15] S. Yang, L. Zhang, L. He, and Y. Wen, "Sparse low-rank component-based representation for face recognition with low-quality images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 251–261, 2019.

[16] J. Li, Y. Kong, H. Zhao, J. Yang, and Y. Fu, "Learning fast low-rank projection for image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4803–4814, 2016.

[17] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 597–604.

[18] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 625–632.

[19] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 676–683.

[20] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1615–1622.

[21] Q. Liu, Z. Lai, Z. Zhou, F. Kuang, and Z. Jin, "A truncated nuclear norm regularization method based on weighted residual error for matrix completion," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 316–330, 2016.

[22] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji, "Discriminant sparse neighborhood preserving embedding for face recognition," *Pattern Recognition*, vol. 45, no. 8, pp. 2884–2893, 2012.

[23] J. Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen, and Y. Xu, "Robust sparse linear discriminant analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[24] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[25] X. Fang, S. Teng, Z. Lai, Z. He, S. Xie, and W. K. Wong, "Robust latent subspace learning for image classification," *IEEE Trans. Neural Netw. Learn. Syst*, 2017.

[26] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[27] P. Zhou, Z. Lin, and C. Zhang, "Integrated low-rank-based discriminative feature learning for recognition," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 5, pp. 1080–1093, 2016.

[28] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, "Collaborative representation based classification for face recognition," *arXiv preprint arXiv:1204.2358*, 2012.

[29] X. Fang, N. Han, J. Wu, Y. Xu, J. Yang, W. K. Wong, and X. Li, "Approximate low-rank projection learning for feature extraction," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[30] W. K. Wong, Z. Lai, J. Wen, X. Fang, and Y. Lu, "Low rank embedding for robust image feature extraction," *IEEE Trans Image Process*, vol. PP, no. 99, pp. 1–1, 2017.

[31] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[32] D. Han and J. Kim, "Unsupervised simultaneous orthogonal basis clustering feature selection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5016–5023.

[33] X. Shi, F. Nie, Z. Lai, and Z. Guo, "Robust principal component analysis via optimal mean by joint l2,1 and schatten p-norms minimization," *Neurocomputing*, vol. 283, pp. 205–213, 2018.

[34] H. Foroughi, N. Ray, and H. Zhang, "Object classification with joint projection and low-rank dictionary learning," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 806–821, 2018.

[35] H. Liu, Y. Liu, Y. Yu, and F. Sun, "Diversified key-frame selection using structured l2,1 optimization," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 3, pp. 1736–1745, 2014.

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[37] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

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

[39] H. Park, "A parallel algorithm for the unbalanced orthogonal procrustes problem," *Parallel Computing*, vol. 17, no. 8, pp. 913–923, 1991.

[40] Z. Ren, B. Wu, X. Zhang, and Q. Sun, "Image set classification using candidate sets selection and improved reverse training," *Neurocomputing*, vol. 341, p. 60–69, 2019.

[41] B. Wang, Q. Yin, S. Wu, L. Wang, and G. Liu, "Discriminative representative selection via structure sparsity," in *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1401–1406.

[42] A. Jain, B. Klare, and A. Ross, "Guidelines for best practices in biometrics research," in *2015 International Conference on Biometrics (ICB)*. IEEE, 2015, pp. 541–545.

[43] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 7, pp. 1490–1507, 2017.

[44] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 663–670.