PLOS ONE

RESEARCH ARTICLE

Subspace structural constraint-based discriminative feature learning via nonnegative low rank representation

Ao Li[1]*, Xin Liu[1], Yanbing Wang[2], Deyun Chen[1], Kezheng Lin[1], Guanglu Sun[1], Hailong Jiang[3]

1 Postdoctoral Station of School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China, **2** School of Measurement–Control Technology and Communications Engineering, Harbin University of Science and Technology, Harbin, China, **3** Department of Computer Science, Kent State University, Kent, United States of America

* dargonboy@126.com

Citation: Li A, Liu X, Wang Y, Chen D, Lin K, Sun G, et al. (2019) Subspace structural constraint-based discriminative feature learning via nonnegative low rank representation. PLoS ONE 14(5): e0215450. https://doi.org/10.1371/journal.pone.0215450

Editor: Kim Han Thung, University of North Carolina at Chapel Hill, UNITED STATES

Received: October 14, 2018

Accepted: April 2, 2019

Published: May 7, 2019

Data Availability Statement: The relevant dataset is available as Supporting Information and from the following URL: http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html.

Funding: The paper was supported in part by National Natural Science Foundation of China (Grant 61501147), University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (Grant UNPYSCT-2018203), Excellent Youth Scholar Fund of Heilongjiang Province (Grant JJ2019YX0116), Fundamental Research Foundation for University of Heilongjiang

Abstract

Feature subspace learning plays a significant role in pattern recognition, and many efforts have been made to generate increasingly discriminative learning models. Recently, several discriminative feature learning methods based on a representation model have been proposed, which have not only attracted considerable attention but also achieved success in practical applications. Nevertheless, these methods for constructing the learning model simply depend on the class labels of the training instances and fail to consider the essential subspace structural information hidden in them. In this paper, we propose a robust feature subspace learning approach based on a low-rank representation. In our approach, the low-rank representation coefficients are considered as weights to construct the constraint item for feature learning, which can introduce a subspace structural similarity constraint in the proposed learning model for facilitating data adaptation and robustness. Moreover, by placing the subspace learning and low-rank representation into a unified framework, they can benefit each other during the iteration process to realize an overall optimum. To achieve extra discrimination, linear regression is also incorporated into our model to enforce the projection features around and close to their label-based centers. Furthermore, an iterative numerical scheme is designed to solve our proposed objective function and ensure convergence. Extensive experimental results obtained using several public image datasets demonstrate the advantages and effectiveness of our novel approach compared with those of the existing methods.

Introduction

Feature subspace learning is a critical technique for feature extraction, which has been widely and well studied in the areas of computer vision, data mining, and pattern recognition [1, 2, 3]. Many representative works have been proposed for feature subspace learning. For example, principal component analysis (PCA) [4] is a classical unsupervised feature learning method, which seeks a subspace with maximum variance of the projected samples to project the high-dimensional data onto a lower dimensional subspace. Aiming to preserve the local

Province (Grant LGYC2018JQ013) and China Postdoctoral Science Foundation (Grant 2016M601438). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. All funding were received during this study.

Competing interests: The authors have declared that no competing interests exist.

neighborhood structure in the data manifold, He *et al*. proposed a neighbor-preserving embedding (NPE) [5], which showed advantages over PCA in terms of the robustness to noise and reduced sensitivity to outliers. Locality-preserving projection (LPP) [6] is another effective feature projection method, which attempts to preserve more local structure of the original image space. Although both NPE and LPP are unsupervised feature learning methods, they can be extended to supervised scenarios to achieve improved performance. To improve the robustness and discriminative ability of preserving projection methods, structurally incoherent low-rank 2DLPP (SILR-2DLPP) [7] was proposed, which realized the discriminability of the preserving projection learning by recovering the sample from different classes. Linear discriminant analysis (LDA) [8] is a well-known supervised subspace learning method, which obtains the projection by Fisher's LDA and produces well-separated classes in a low-dimensional subspace with discriminative information. For further improvement, locality-sensitive discriminant analysis (LSDA) [9] was presented, which aimed to learn projection by determining the local manifold structure to maximize the margin between the data points from the different classes in each local area.

Recently, several feature extraction techniques based on representation models have received increased attention owing to their robustness. Among them, sparse representation (SR) and low-rank representation (LRR) are two representative models that are the most well-known and widely used in many recognition and classification applications. Wright *et al*. proposed a SR-based classification (SRC) method [10]. In [10], SR was used to represent the test sample with the smallest number of training instances, and the representation coefficients were considered as its features to determine the classification results of the test sample. It was verified through a facial recognition problem that SRC provided excellent and robust results despite the facial occlusions. To reveal the essential mechanism of SRC, a collaborative representation providing an interesting analysis of the representation-based facial recognition framework was proposed to extract the coding features [11]. Many existing SRC-based coding schemes lead to significant classification errors because they ignore the relevance between similar instances. In [12], Li *et al*. proposed a self-supervised sparse coding scheme for image classification based on LRR, which could effectively preserve the local structure information of the coding for similar instances. Moreover, the main concept of SRC has also been extended to applications in subspace learning. In [13], Zhang *et al*. proposed a novel linear subspace learning approach by combining sparse coding and feature grouping. In their method, a dictionary was learned from the training dataset and used to sparsely decompose the training samples. Then, the decomposition components were divided into more and less discriminative parts, respectively, to learn the desired subspace. However, when the training instances are corrupted, this method is not sufficiently robust to the noise. By giving an interpretation from a probabilistic view, Cai *et al*. proposed a probabilistic collaborative representation-based classification (ProCRC) model in which the probability of a test belonging to the collaborative subspaces of all the classes was well defined by the learned coding feature [14]. They also reported that ProCRC achieved good results for numerous pattern classification problems.

To increase the robustness, low-rank models have attracted significant attention owing to their effectiveness in recovering data and removing noise. They have already been applied to many fields including dictionary learning [15, 16], transfer learning [17], domain adaptation [18, 19], and outlier detection [20]. As an extension of SR, LRR not only solves the subspace recovery problem but also captures the low-dimensional subspace structures accurately [21]. Subsequently, many LRR-based feature learning methods have been studied in recent years. Liu *et al*. proposed a latent LRR model to integrate subspace segmentation and feature extraction in a unified framework, which could robustly extract the salient features from the data by exploiting the latent structural information hidden in the data [22]. Considering the drawbacks

of learning two low-rank matrices individually in latent LRR, a supervised feature extraction method by approximating the LRR was proposed by Fang *et al.* [23]. Unlike latent LRR, they treated the above-mentioned two matrices as a combined matrix during learning, which mutually boosted them and extracted more discriminative features. Zhang *et al.* proposed a structural LRR [24]. In [24], ideal supervised regularization was introduced to guide the feature learning process and low-rank recovery was performed for the training data from all the classes simultaneously without losing the structural information. The results showed that the obtained features consisting of the representation coefficients were suitable for classification. Ma *et al.* proposed a discriminative low-rank dictionary learning algorithm for sparse representation (DLRDSR) [25], which combined low-rank constraints and discriminative dictionary learning to perform SR for solving the problem of facial recognition. Zhou *et al.* integrated latent LRR with a ridge regression-based classifier, which could place feature learning and classification in the same framework. Consequently, the classifier and feature learning could benefit each other during the iteration, and the learned feature was more adaptive to the classification problem [26]. To increase the discrimination, Luo *et al.* proposed feature learning with calibrated data reconstruction and a low-rank model. By minimizing the joint $l_{2,1}$-norm reconstruction error and inner-class distance, the discriminative information and reconstructed low-rank structures were preserved simultaneously, which helped improve the feature learning [27].

Motivated by the success of representation-based feature learning, this paper proposes a nonnegative LRR-based robust and discriminative feature learning method for image classification, in which the LRR and feature subspace learning are combined in a unified framework. In our proposed framework, the nonnegative LRR coefficients, as the measurements of the low-dimensional structural similarity, are utilized to guide the feature subspace learning. Thus, the LRR coefficients are introduced as weighted constraints on the distances of the pairs of the projected instances in the feature subspace. In addition, the feature subspace learning and LRR can benefit from each other during the iteration to ensure an overall optimum. Furthermore, to address the classification problem, we incorporate a discriminative linear regression term in the proposed framework, which can be used to provide an additional supervised effect. Thus, it will enable our model to learn a more discriminative feature subspace and be more adaptive to the classification task. Extensive experiments are conducted on several public datasets and encouraging results are obtained.

The contributions of our work are as follows:

1. We design a new feature learning model that incorporates LRR into feature subspace learning. In our proposed model, the LRR coefficients are exploited as the similarity measurements to guide the feature learning dynamically and adaptively. Furthermore, a class-label-based linear regression is incorporated into the proposed model as extra supervised information to further improve the performance, which can make the extracted features to be more discriminative and adaptive for classification tasks.
2. We introduce a nonnegative constraint to the LRR coefficients in our proposed objective function. The coefficients can be used as penalty parameters for penalizing the approximation of the related instances in the feature subspace, which will adaptively lead to a small inner-class with a large intra-class margin.
3. We develop an iterative scheme with the recent augmented Lagrangian multiplier (ALM) method [28] and Block coordinate descent(BCD)[29] in which the objective function is solved efficiently and convergence is ensured.
4. We evaluate our approach using several image datasets with different classifiers to show the effectiveness and robustness of our novel model.

The remainder of this paper is organized as follows. The related works on LRR and discriminative feature learning are reviewed in the second section. The third section elaborates our proposed approach followed by the theoretical analysis and development of the numerical scheme. The experimental and analysis results are reported in the fourth section. The fifth section concludes this paper.

Related work

In this section, we briefly review the related works on LRR and discriminative subspace learning, respectively.

Low-rank representation

LRR has drawn great attention and has already been applied in many fields such as subspace learning [30, 31], subspace clustering [21, 32], and image processing [33, 34]. Consider a set of data samples $X \in R^{m \times n}$ (m and n denote the dimension and number of samples, respectively), which can be represented by the linear combination of the basis in dictionary A and error components E. The object function of LRR is as follows:

$$\min_{Z,E} rank(Z) + \lambda \|E\|_l, \; s.t. X = AZ + E \tag{1}$$

where Z denotes the representation coefficient matrix, $rank(\cdot)$ denotes the rank of the matrix, and $\|\cdot\|_l$ indicates the norm-based regularization strategy applied to the error matrix, such as Frobenius norm, l_1 norm, or $l_{2,1}$ norm. All the three norms can be used to model the corruption and outlier existing in the data. Nevertheless, the $l_{2,1}$ norm shows some advantages in exploring the relevance in the data and can well characterize the sample-specific corruption. λ is a penalty parameter for balancing the two terms.

However, the rank-minimization problem expressed in Eq (1) is difficult to solve because the rank function is nonconvex. To address this problem, the nuclear norm, which is a convex relaxation of the rank operator [22, 35], is used to replace the first term in Eq (1). Hence, the object function can be rewritten as follows:

$$\min_{Z,E} \|Z\|_* + \lambda \|E\|_l, \; s.t. X = AZ + E \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm of matrix that computes the sum of singular values of the matrix [36]. If we take the data matrix itself as dictionary A, then Eq (2) is converted into the following self-expression form:

$$\min_{Z,E} \|Z\|_* + \lambda \|E\|_l, \; s.t. X = XZ + E \tag{3}$$

It is reported that the representation coefficients in Eq (3) can well present the similarity in the manifold structure of the instances themselves to some extent [37]. Based on this assumption, a graph learning model was constructed with LRR in [37] for the clustering and recognition issues. Motivated by [37], we also wanted to incorporate LRR into the feature subspace learning and consider the coefficient as the similarity measurement to constrain the distance of the feature subspace of the instances.

Feature subspace learning

Lately, feature subspace learning is becoming well known and practical, and it can be divided into three categories: unsupervised methods, supervised methods, and semi-supervised methods. The concept of subspace learning is learning a projection subspace that can project high-

dimensional data onto a low-dimensional space [38]. Concurrently, useful information is retained, and the similarity of the inner-class and dissimilarity of the inter-class can be further increased. To this end, discriminative feature learning methods are well studied and have recently become a very active topic.

LDA is one of the most common supervised subspace learning methods aimed at finding a projection that maximizes inter-class scatter and minimizes intra-class scatter simultaneously. Supervised subspace learning methods can effectively extract discriminative information and achieve improved classification performance.

Considering a training dataset X with multiclass instances, the inter-class divergence can be formulated as follows:

$$S_b = \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T \tag{4}$$

where C denotes the number of classes, μ denotes the mean vector of the whole training dataset, and μ_i denotes the mean vector of training instances that belongs to the i-th class.

Similarly, the summation of intra-class divergence for the training data can be formulated as follows:

$$S_w = \sum_{i=1}^{C} \sum_{j=1}^{M_i} (x_i^j - \mu_i)(x_i^j - \mu_i)^T \tag{5}$$

where M_i indicates the number of training samples within the i-th class, and x_i^j represents the j-th instance that belongs to the i-th class.

After defining the above two kinds of divergence, the objective function for LDA is formulated as the following maximum problem:

$$\max_P \frac{tr(P^T S_b P)}{tr(P^T S_w P)} \tag{6}$$

where P denotes the feature projection subspace to be learned.

Using the Lagrangian multiplier method with λ, the above Eq can be transformed into a problem of solving eigenvectors as follows:

$$S_b P = \lambda S_w P \tag{7}$$

With Eq (7), LDA learns a discriminative subspace that consists of the first C–1 eigenvectors of matrix $S_w^{-1} S_b$.

To improve the performance of LDA, many extended methods based on it have been proposed in recent years. Local Fisher discriminant analysis (FDA) [39] is a new linear supervised dimensionality reduction method that effectively combines the concepts of FDA [40] and LPP [41]. Subsequently, semi-supervised local FDA was proposed to preserve the global structure of the unlabeled samples in addition to separating the labeled samples into different classes [42]. Probabilistic LDA (PLDA) is a generative probability model that can extract and combine features for recognition [43]. With PLDA, a model of a previously unseen class can be built from a single example and multiple examples can be combined for improving the representation of the class. Sparse discriminant analysis (SDA) is a method for performing LDA under an imposed sparseness criterion [44]. SDA is based on the optimal scoring interpretation of LDA and can be extended to perform sparse discrimination by mixtures of Gaussians if the boundaries between the classes are nonlinear or if the subgroups are presented within each class.

Motivated by the above insights, we want to incorporate an LRR into the feature subspace learning and to consider the class label and LRR coefficient as two different types of constraint items to maximize the inter-class scatter and minimize the intra-class scatter simultaneously. To this end, a structural similarity-based constraint term is designed by first utilizing the LRR coefficients. Next, a label-based linear regression constraint is incorporated to achieve extra discrimination and adaptation to the classification problem.

Our proposed approach

In this section, our discriminative feature learning model is proposed and the novel objective function for our proposed model is detailed and analyzed. To solve the objective function efficiently, we also developed a numerical scheme to obtain an approximate solution.

Construction of proposed feature subspace learning

As mentioned above, in conventional LDA-based approaches, the constraint term for the feature subspace learning is combined with the label information, aiming to enforce the minimum intra-class distance and maximum inter-class distance within the learned subspace. This can be considered as a learning strategy with equal weights assigned to different training instances. In such methods the regularization parameters are 1 for the pairs of training instances within the same class and -1 for those belonging to different classes. However, the equal-weighted scenario is not typically optimal. On one hand, the differences between the instances from the same class may not be uniformly closed owing to some objective factors. For example, face instances can suffer from expression or lighting variation. On the other hand, in the real world, the data are contaminated with noise and outliers, which can disrupt the essential structural relevance. However, it has been found that instances from same class generally lie in the same low-dimensional subspace [10]. Furthermore, the essential structural information can be explored with an LRR model even when the data are corrupted.

Based on the above observations, our basic concept is to introduce low-dimensional structural information in the constraint on the feature subspace, which can lead to learning a robust and adaptive feature subspace. Thus, our feature learning objective function can be defined as follows:

$$\begin{gathered}\min_{P,Z,E}\eta\|Z\|_{*}+\sum_{ij}Z_{ij}\|P^{T}X_{i}-P^{T}X_{j}\|_{2}^{2}+\lambda\|E\|_{2,1}\\ s.t.X=XZ+E,\ Z_{ij}\geq 0\end{gathered}\tag{8}$$

where $X = [X_1,X_2,\ldots,X_m]$ is the training set, $X_i(i = 1,2,\ldots,m)$ represents each column of X, and m is the total number of training instances. P and E denote the feature subspace and error matrix, respectively. η and λ are positive scalars to balance the three terms. The first term in Eq (8) is used to enforce a low-rank constraint on representation matrix Z, which helps explore the low-dimensional structures hidden in the training instance. The second term is our proposed constraint for the feature subspace, which considers the low-rank coefficients as the similarity weights for constraining the distances of the pairs of projected instances. It is noted that each element of Z can be considered as a measurement of the low-dimensional structural similarity for each pair of instances. Thus, using our proposed constraint term, the structural similarity information is not only preserved in the learned subspace but also used to guide the feature learning. In addition, with the second term, Z and P can be learned jointly, benefiting each other during the iteration and yielding progressively better and robust solutions. Moreover, we also introduce a nonnegative constraint in each element of Z, which can ensure that Z

is used as a nonnegative regularization parameter. To increase the robustness, the third term enforces the $l_{2,1}$ norm-based constraint on the error matrix, which is used to better explore the relevance in the data and combat sample-specific corruptions.

As described in [21], the LRR matrix can lead to large values for instances lying in the same low-dimensional subspace and small values for those in different subspaces. In addition, closeness of the two instances implies a large Z_{ij} and vice versa. Hence, different from the conventionally designed feature learning, our feature learning constraint can effectively optimize both the intra-class and inter-class divergences with some adaptive structural similarity information from the latent low-dimensional space. To avoid trivial solutions and reduce the redundancy, an orthogonal constraint is also imposed on feature subspace P. Thus, the minimization problem in (8) can be rewritten as

$$\begin{aligned} &\min_{P,Z,E} \eta\|Z\|_* + \sum_{ij} Z_{ij}\|P^T X_i - P^T X_j\|_2^2 + \lambda\|E\|_{2,1} \\ &s.t. X = XZ + E,\ Z_{ij} \geq 0,\ P^T P = I \end{aligned} \tag{9}$$

where I is the identity matrix.

To make our model more discriminative and adaptive in the classification task, the label information was incorporated into our framework as a kind of discriminative supervised information. To this end, the comprehensive objective function for our proposed framework is reformulated as follows:

$$\begin{aligned} &\min_{P,Z,E} \frac{1}{2}\|Y - P^T X\|_F^2 + \sum_{ij} Z_{ij}\|P^T X_i - P^T X_j\|_2^2 + \eta\|Z\|_* + \lambda\|E\|_{2,1} \\ &s.t. X = XZ + E,\ Z_{ij} \geq 0,\ P^T P = I \end{aligned} \tag{10}$$

where $Y = [Y_1, Y_2, \ldots, Y_m]$ is a matrix decided by the class label. $Y_i = [-1, -1, \ldots, 1, \ldots, -1]^T \in R^C$ denotes the i-th column of Y, and its c-th element is 1, whereas the others are -1 if the i-th instance belongs to the c-th class. With the first label fitness term in Eq (10), our feature subspace will be jointly learned by minimizing the classification error simultaneously.

In our proposed framework, the first two terms in Eq (10) can be considered as two types of effective constraints for optimizing the learned feature subspace. From Fig 1(A), we can see that with the first term, the class label can be used to provide a clustering center, which will enable the learned subspace to be discriminative and adaptative for the classification problem. However, the inferior inter-class divergence due to the corruption still needs to be optimized, as shown in Fig 1(A). To this end, we incorporated the second term into the framework. Consequently, the feature learning is guided by the low-dimensional adaptive structural information, which increases the small intra-class divergence and large inter-class divergence within the projected feature subspace as shown in Fig 1(B) and helps to improve the learning performance.

Solution scheme for our novel objective function

In this section, we develop an iterative numerical scheme for solving the novel objective function. It is worth noting that the minimization problem in Eq (10) is not jointly convex with respect to all the variables [45]. Hence, the inexact ALM with BCD are used to obtain the approximate solution. To decouple the variables, two auxiliary variables W and M are

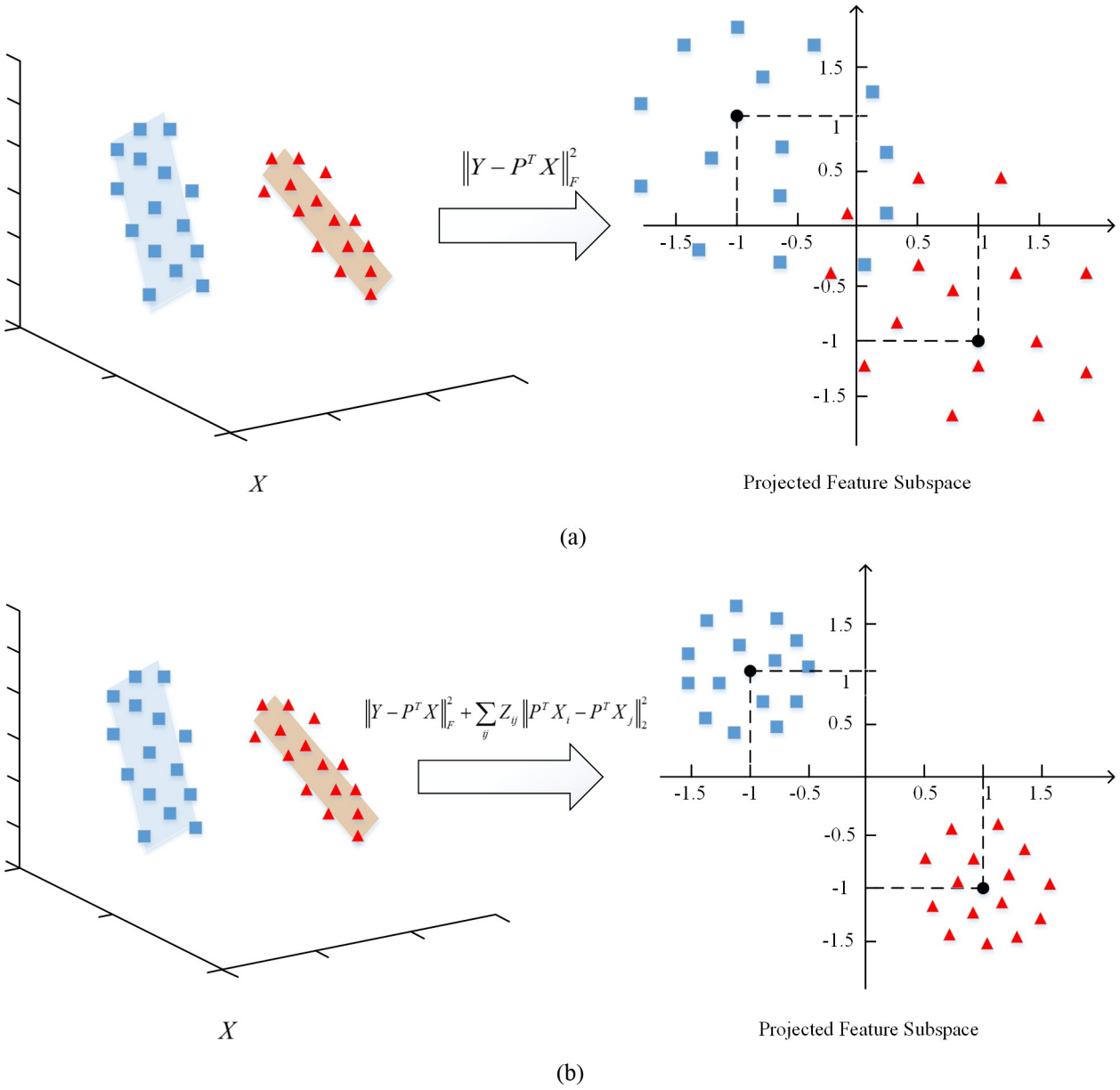


Fig 1. Graphical analysis of our proposed framework. (a) the effect of the first constraint term and (b) the effect of the incorporated first two constraint terms.

https://doi.org/10.1371/journal.pone.0215450.g001

introduced to relax the minimization, and the objective function can be rewritten as follows:

$$\min_{P,Z,E,W,M} \frac{1}{2}\|Y - P^T X\|_F^2 + \sum_{ij} M_{ij}\|P^T X_i - P^T X_j\|_2^2 + \eta\|W\|_* + \lambda\|E\|_{2,1}$$
$$s.t. X = XZ + E, Z = W, Z = M, M_{ij} \geq 0, P^T P = I \tag{11}$$

By the ALM, the Lagrangian function of problem in Eq(11) is

$$\begin{aligned}\mathcal{L}(P,Z,E,W,M) = &\frac{1}{2}\|Y-P^TX\|_F^2+\sum_{ij}M_{ij}\|P^TX_i-P^TX_j\|_2^2+\eta\|W\|_*+\lambda\|E\|_{2,1}\\&+\frac{\mu}{2}(\|X-XZ-E\|_F^2+\|Z-W\|_F^2+\|Z-M\|_F^2)\\&+\langle Y_1,X-XZ-E\rangle+\langle Y_2,Z-W\rangle+\langle Y_3,Z-M\rangle\end{aligned} \tag{12}$$

where $\langle\bullet\rangle$ denotes the operation for inner product and $Y_i(i = 1,2,3)$ is the Lagrangian multiplier. Then, we transform $\mathcal{L}$ into the following compact form.

$$\begin{aligned}\mathcal{L}(P,Z,E,W,M) = &\frac{1}{2}\|Y-P^TX\|_F^2+\sum_{ij}M_{ij}\|P^TX_i-P^TX_j\|_2^2+\eta\|W\|_*+\lambda\|E\|_{2,1}\\&+\frac{\mu}{2}\left(\left\|X-XZ-E-\frac{Y_1}{\mu}\right\|_F^2+\|Z-W-\frac{Y_2}{\mu}\|_F^2+\left\|Z-M-\frac{Y_3}{\mu}\right\|_F^2\right)\\&-\frac{1}{2\mu}(\|Y_1\|_F^2+\|Y_2\|_F^2+\|Y_3\|_F^2)\end{aligned} \tag{13}$$

Thus, the minimization can be converted as

$$\begin{aligned}&\min_{P,Z,E,W,M}\mathcal{L}(P,Z,E,W,M)\\&s.t.M_{ij}\geq 0,P^TP=I\end{aligned} \tag{14}$$

With the recently proposed BCD, the minimization can be solved iteratively for each variable while others are fixed. In this way, in the k-th iteration, the projection subspace P can be learned as

$$\begin{aligned}&\min_P\frac{1}{2}\|Y-P^TX\|_F^2+\sum_{ij}M_{ij}^k\|P^TX_i-P^TX_j\|_2^2\\&s.t.P^TP=I\end{aligned} \tag{15}$$

To solve Eq (15) efficiently, we first rewrite it as the following graph-based compact formulation:

$$\begin{aligned}&\min_P\frac{1}{2}\|Y-P^TX\|_F^2+Tr(P^TXLX^TP)\\&s.t.P^TP=I\end{aligned} \tag{16}$$

where $L = D–M$ denotes the graph Laplacian matrix, and D presents a diagonal matrix with $D_{ii}=\frac{\sum M_{*i}+\sum M_{i*}}{2}$. Owing to the orthogonal constraint, the minimization cannot be considered as an easy quadratic problem. Given the derivative of the objective function in Eq (16) as follows, it can be solved with the method proposed in [46].

$$\frac{\partial\mathcal{L}_P}{\partial P}=XX^TP-XY^T+XLX^TP \tag{17}$$

Similarly, by fixing other variables, the objective function with respect to W is shown as

$$\min_W\left\|W-\left(Z^k-\frac{Y_2^k}{\mu}\right)\right\|_F^2+\eta\|W\|_* \tag{18}$$

Eq (18) is a classical rank-minimization problem that can be solved efficiently by the singular value shrinkage operator [47].

Next, ignoring the variables independent of Z in Eq(13), we have

$$\min \left\| X - XZ - E^k - \frac{Y_1^k}{\mu} \right\|_F^2 + \left\| Z - W^{k+1} - \frac{Y_2^k}{\mu} \right\|_F^2 + \left\| Z - M^k - \frac{Y_3^k}{\mu} \right\|_F^2 \tag{19}$$

It is worth noting that Eq (19) is a quadratic convex minimization, which can be solved by forcing its derivative to zero. Thus, we can obtain its closed-form solution as

$$Z^{k+1} = (2I + X^T X)^{-1}(W^{k+1} + M^k - X^T E^k + X^T X - (-X^T Y_1^k + Y_2^k + Y_3^k)/\mu) \tag{20}$$

After dropping the terms irrelevant to M, we can obtain

$$\min_M \left\| Z^{k+1} - M - \frac{Y_3^k}{\mu} \right\|_F^2 + \sum_{ij} M_{ij} \| P^{T(k+1)} X_i - P^{T(k+1)} X_j \|_F^2 \tag{21}$$

For clarity, we rewrite it as the following form

$$\begin{gathered} \min_M \left\| M - \left(Z^k - \frac{Y_3^k}{\mu} \right) \right\|_F^2 + \sum_{i,j} (S^k \otimes M)_{ij} \\ s.t. M_{ij} \geq 0 \end{gathered} \tag{22}$$

where S is a matrix with $S_{ij} = \| P^{T(k+1)} X_i - P^{T(k+1)} X_j \|_2^2$. Moreover, because both of S and M are nonnegative, the minimization in Eq (22) can be converted as

$$\begin{gathered} \min_M \left\| M - \left(Z^{k+1} - \frac{Y_3^k}{\mu} \right) \right\|_F^2 + \| S^{k+1} \otimes M \|_1 \\ s.t. M_{ij} \geq 0 \end{gathered} \tag{23}$$

The problem in Eq (23) can be seen as the nonnegative weighted l_1-norm minimization problem, which can be solved using the method in [48].

Then, by fixing others, the error matrix E can be updated as

$$\min_E \frac{\lambda}{\mu} \| E \|_{2,1} + \frac{1}{2} \left\| E - \left(X - XZ^{k+1} + \frac{Y_1^k}{\mu} \right) \right\|_F^2 \tag{24}$$

The minimization in Eq (24) can be easily solved with the method in [49]. By setting $\Phi = X - XZ^{k+1} + \frac{Y_1^k}{\mu}$, the i-th column of updated E^{k+1} is computed as

$$E_i^{k+1} = \begin{cases} \dfrac{\| \Phi_i \|_2 - \lambda}{\| \Phi_i \|_2}, & if\ \lambda < \| \Phi_i \|_2 \\ 0, & otherwise \end{cases} \tag{25}$$

As stated in the inexact ALM algorithm, the Lagrangian multipliers also need to be updated during the iteration. The details of the developed scheme are summarized in **Algorithm 1**.

Algorithm 1 Scheme for discriminative feature subspace learning

Input: training data X, label matrix Y, $Z = W = M = 0$, $E = 0$, $Y_1 = Y_2 = Y_3 = 0$, $\mu = 0.6$, $\mu_{max} = 10^{10}$, $\rho = 1.1$

Output: P

While not convergence **do**

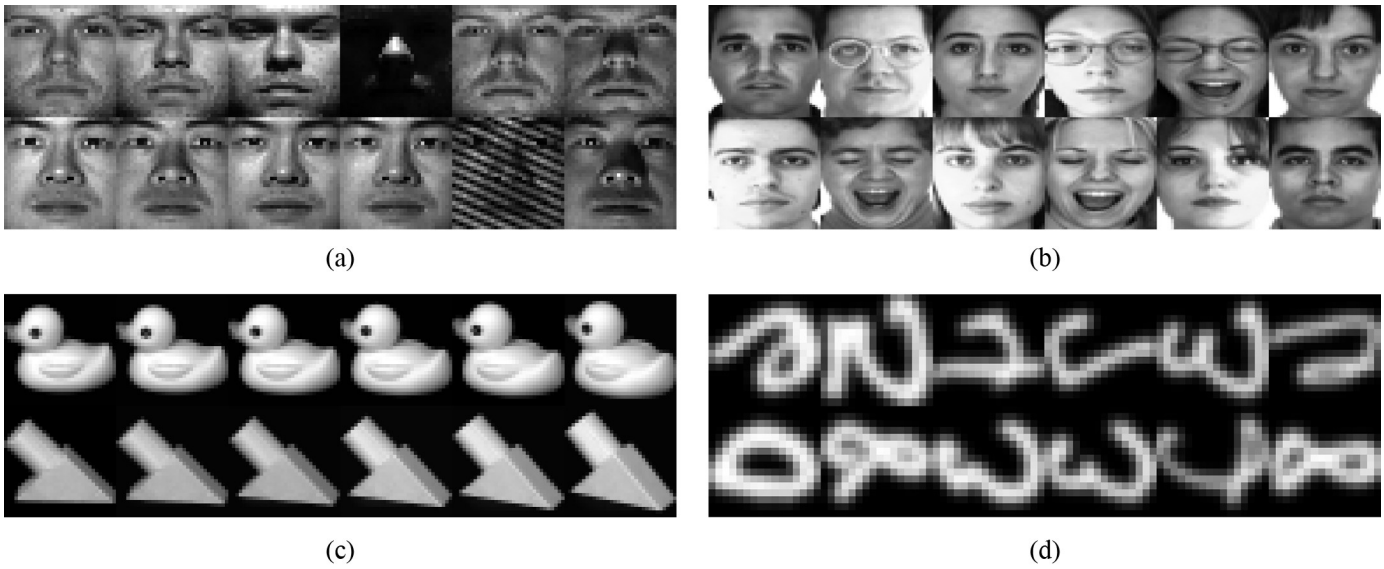


Fig 2. Sample images. (a) Extended YaleB, (b) AR, (c) COIL20, (d) USPS.

https://doi.org/10.1371/journal.pone.0215450.g002

Table 1. Classification rates(%) of comparison methods on test datasets with KNN.

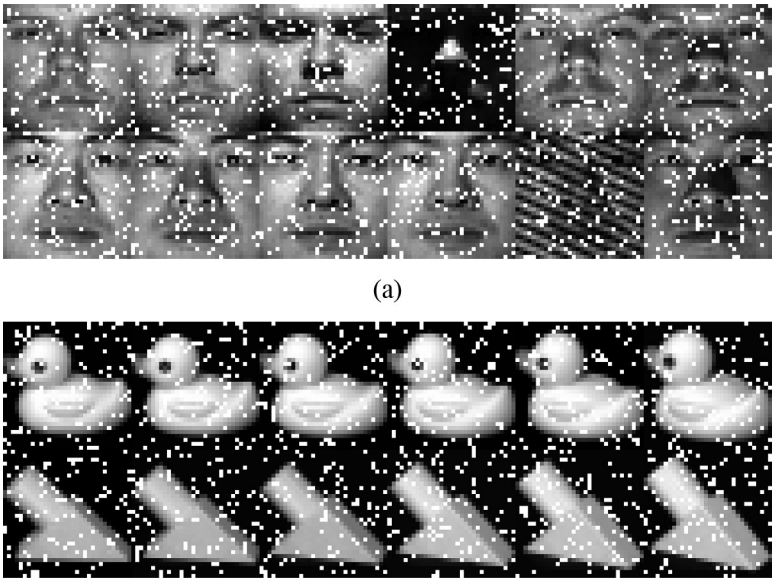
Methods	Extended YaleB	AR	COIL20	USPS
PCA	72.57±0.58	79.43±0.79	89.51±0.67	79.47±1.17
LDA	89.09±0.91	84.68±1.81	89.38±0.84	72.49±0.53
NPE	86.01±1.37	81.26±1.41	85.51±1.26	62.10±2.66
LSDA	92.94±0.88	74.34±0.63	84.23±1.52	56.18±2.67
Latent LRR	88.76±1.26	82.49±3.16	90.08±0.89	81.43±1.39
ProCRC	93.61±0.49	86.86±0.82	84.60±1.71	78.06±2.33
DLRDSR	93.56±1.25	80.52±1.35	88.87±0.93	77.89±1.81
SFE-ALR	92.15±1.33	84.89±0.42	87.12±0.45	77.63±2.66
Ours	95.29±0.43	86.63±1.46	92.03±1.05	81.51±0.82

https://doi.org/10.1371/journal.pone.0215450.t001

Table 2. Classification rates(%) of comparison methods on test datasets with SRC.

Methods	Extended YaleB	AR	COIL20	USPS
PCA	80.29±1.28	81.24±1.13	78.94±1.38	76.10±1.72
LDA	82.58±1.32	93.93±1.30	82.81±0.75	59.12±4.21
NPE	76.85±1.51	81.47±1.08	82.59±1.30	60.70±5.39
LSDA	87.53±1.08	81.54±1.26	61.03±4.72	76.14±2.53
Latent LRR	94.37±1.36	95.14±1.64	87.25±3.05	78.91±0.91
ProCRC	93.87±1.83	93.92±0.61	86.45±1.68	77.35±0.86
DLRDSR	92.66±1.74	90.37±1.16	86.53±1.47	77.43±1.31
SFE-ALR	92.70±0.87	95.43±0.67	85.81±0.93	77.97±0.96
Ours	95.86±0.34	96.92±0.94	88.97±1.18	79.75±0.83

https://doi.org/10.1371/journal.pone.0215450.t002



(a)

(b)

Fig 3. Sample images with 10% noise. (a) Extended YaleB, (b) COIL20.

https://doi.org/10.1371/journal.pone.0215450.g003

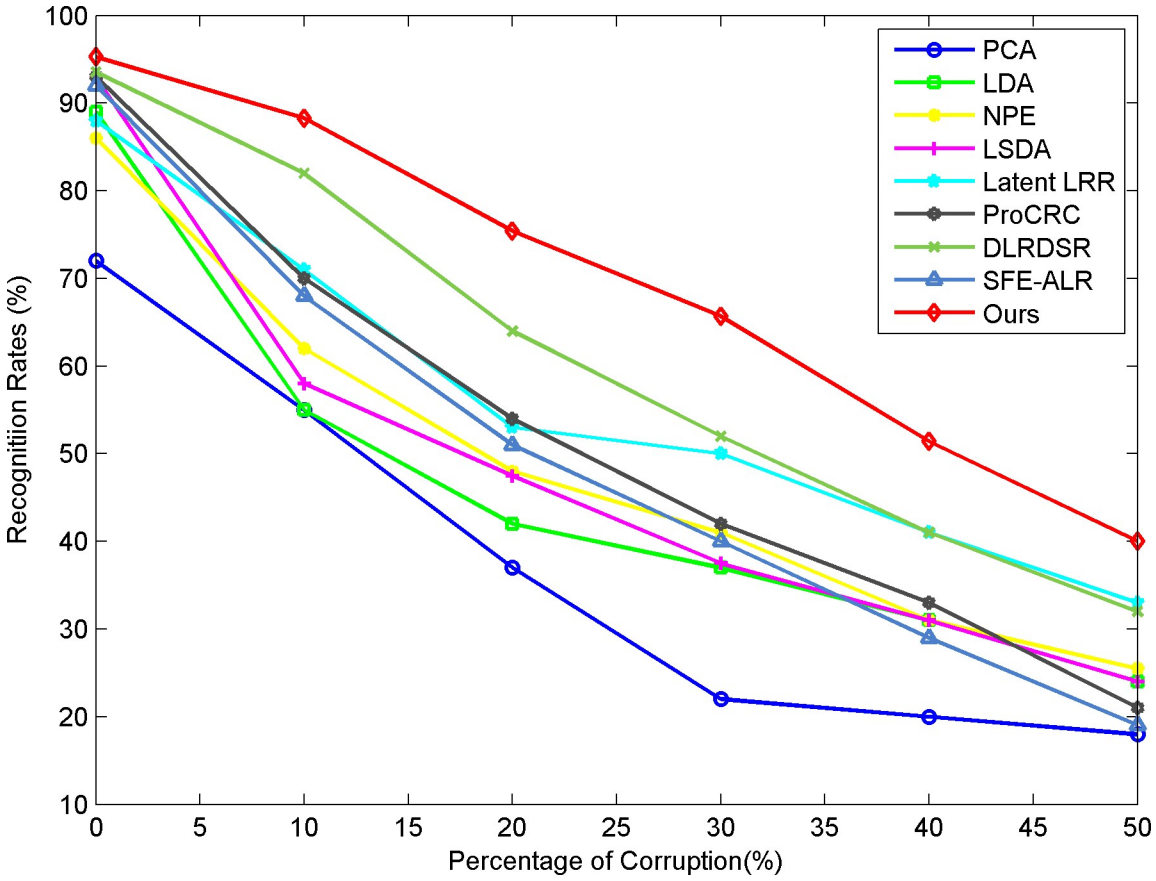


Fig 4. Recognition results versus pixel corruption on extended YaleB.

https://doi.org/10.1371/journal.pone.0215450.g004

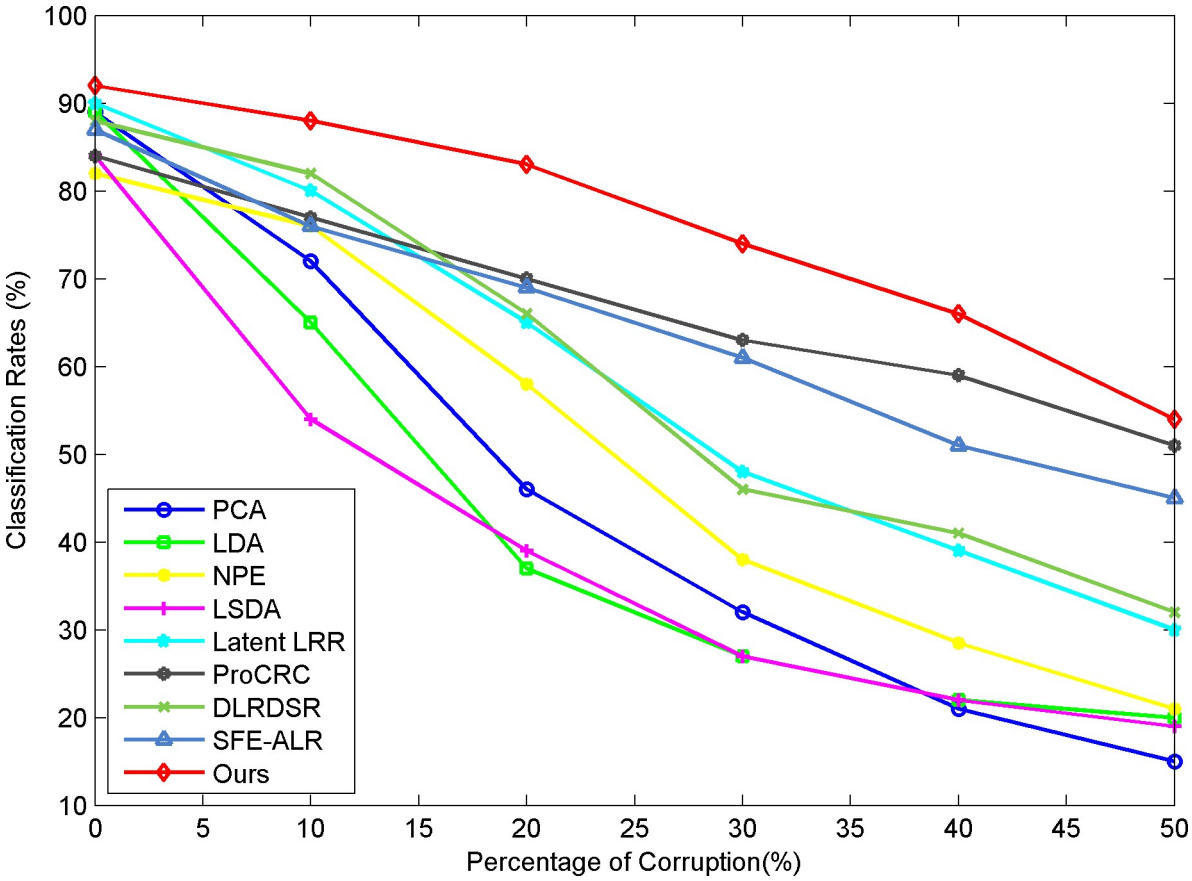


Fig 5. Classification results versus pixel corruption on COIL20.

https://doi.org/10.1371/journal.pone.0215450.g005

1. Update P^{k+1} using Eq (16)
2. Update W^{k+1} using Eq (18);
3. Update Z^{k+1} using Eq (20);
4. Update M^{k+1} using Eq (23);
5. Update E^{k+1} using Eq (24);
6. Update the Lagrangian multipliers and parameter:
   $Y_1^{k+1} = Y_1^k + \mu(X - XZ^{k+1} - E^{k+1})$
   $Y_2^{k+1} = Y_2^k + \mu(Z^{k+1} - W^{k+1})$
   $Y_3^{k+1} = Y_3^k + \mu(Z^{k+1} - W^{k+1})$
   μ = min($\mu_{\max}$, $\rho\mu$);

end while

With the numerical scheme in Algorithm 1, optimal feature subspace P^* can be learned when it achieves convergence. Subsequently, the feature can be extracted by projecting each sample x onto P^* as P^*x, and classification or recognition methods can be implemented on the projected features.

In Algorithm 1, Steps 1 to 4 will consume the most time. The computation complexity with respect to both P and W is $O(n^3)$ owing to the singular value decomposition. The computational cost of solving Z is approximately $O(n^3)$, which is equivalently a matrix inverse calculation. For M, it can be seen as a nonnegative weighted l_1-norm minimization problem, and its complexity is $O(n^2)$. Therefore, the total computation complexity of Algorithm 1 is $O(tn^3)$, where t is the number of iterations.

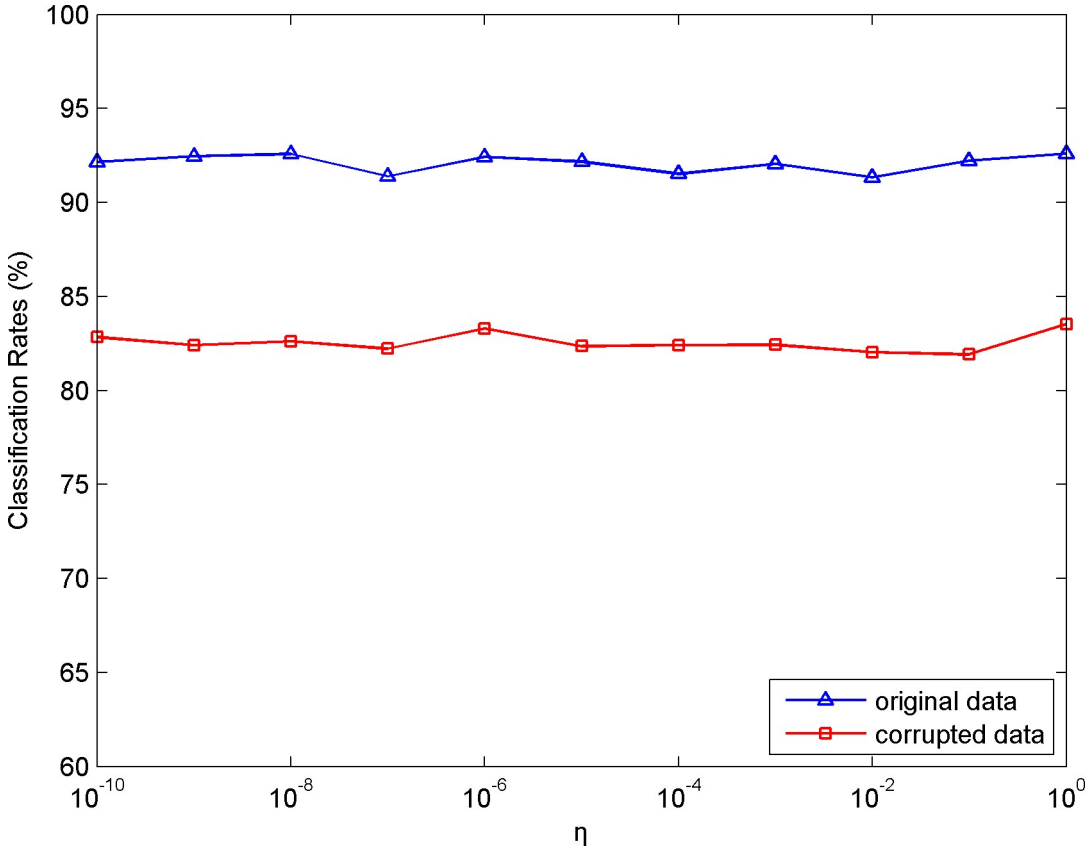


Fig 6. Classification results versus variational η.

https://doi.org/10.1371/journal.pone.0215450.g006

Experimental results and discussion

Experimental results

In this section, we evaluate our proposed approach using four available public datasets. The public datasets include two face datasets: object dataset and handwriting dataset. The details of the datasets are described below. For the face datasets, the individual in this manuscript has given written informed consent to publish these case details.

Extended YaleB dataset. The Extended YaleB face dataset includes 2414 frontal images of 38 individuals, and of each individual, there are approximately 64 images under different lighting conditions. Some instance images are shown in Fig 2(A). The sizes of the test images used in our experiment are cropped to be 32×32. In addition, the samples are normalized to have a unit norm. Thirty two images of each individual are randomly selected as the training set, whereas the remaining are used as the test set.

AR dataset. The AR dataset has 3120 gray images of 120 individuals. For each individual, there are 26 images from the frontal views with different expressions, lighting conditions, and occlusions. Some samples are shown in Fig 2(B). In our experiments, all the face images are cropped and then resized to 55×40. Half of the images of each individual are used for training, and the remaining are used for testing.

Object dataset. COIL20 contains of 1440 images from 20 objects, and each object has 72 images captured from continuous angles at intervals of 5 degree, as shown in Fig 2(C). In our experiment, all the images in dataset are resized to 32×32 and normalized. Ten images of each object are used for training, and the remaining are used for testing.

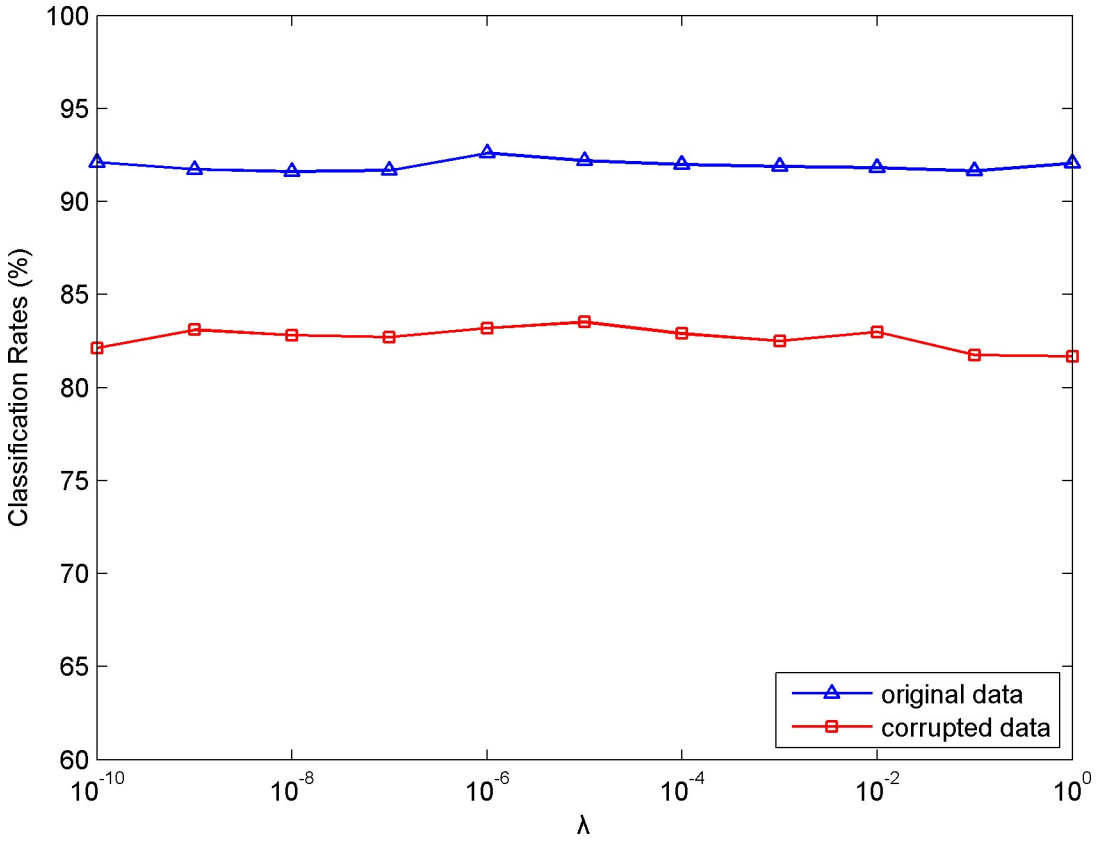


Fig 7. Classification results versus variational λ.

https://doi.org/10.1371/journal.pone.0215450.g007

Handwritten dataset. The USPS dataset has totally 9298 handwritten digit images with ten classes from zero to nine, of which some instances are shown in Fig 2(D). The size of each image is 16×16. In the experiment, for each digit, we randomly select 10 images to group the training set, and the remaining ones are used for testing.

In our experiments, we compared the proposed approach to several existing excellent methods for feature subspace learning, including PCA, LDA, NPE, LSDA, latent LRR in [22], ProCRC [14], DLRDSR [25], and SFE-ALR [23] respectively. Without loss of generality, we use two types of classifiers, SRC and KNN, to test the comparison methods on the test datasets. For SRC, the training instances are used as the atoms in the dictionary, and the recognition or classification results are decided by the minimum class-specific regression error. For KNN, the classification results are decided by the first K neighbors within the feature subspace, and K is set as 1 in our experiments. All the experiments for each dataset are implemented five times. The average classification results with KNN and SRC are reported with the standard deviations in Table 1 and Table 2, respectively.

It can be seen from Tables 1 and 2 that our proposed approach shows a better performance than the other comparison methods on practically all the testing datasets. Moreover, the advantages are obtained consistently with both the KNN and SRC classifiers, implying that the proposed approach exhibits a stable performance compared to that of the classification models. The reasons for our better performance are that the underlying subspace structure is well studied with the low-rank model and its coefficients are effectively used as the relevance measurements to constrain the learned projection. Moreover, by incorporating an LRR into the

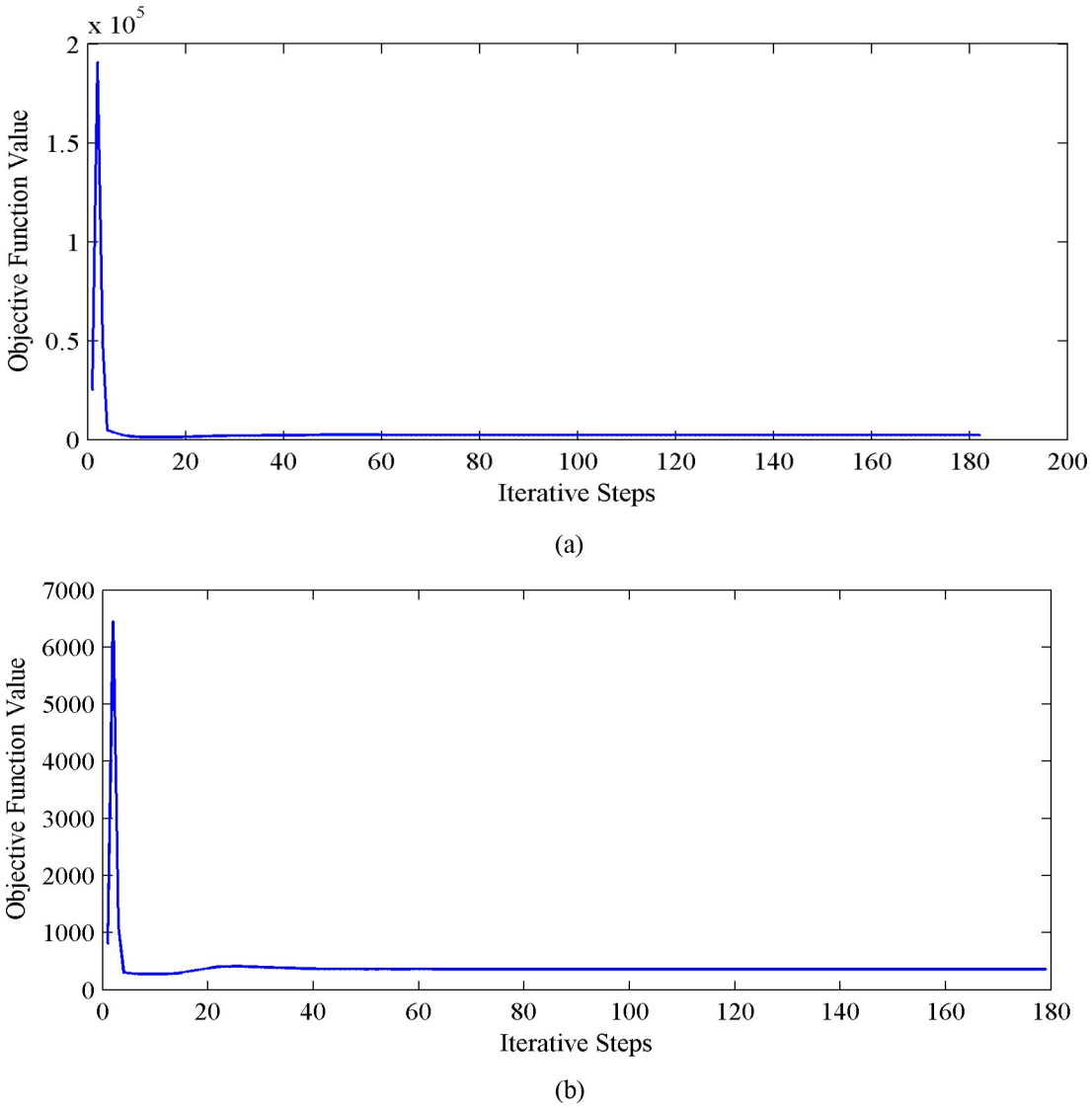


Fig 8. Objective function values versus iterative steps. (a) curve for Extended YaleB dataset and (b) curve for COIL20 dataset.

https://doi.org/10.1371/journal.pone.0215450.g008

feature subspace learning, these two variables can benefit each other during the iteration and obtain jointly optimal solutions.

To test the robustness of our approach, we add random impulse noises of different levels to two selected datasets, Extended YaleB and COIL20. The pixels are corrupted with different percentages of the original image, and some instances of the noisy images can be found in Fig 3. The classification results are shown in Fig 4 and Fig 5. The results are obtained under the same parameter settings for the experiments on clean datasets. From the classification results, we can see that the LRR-based methods show advantages under the noisy conditions compared to the conventional methods for feature learning. This is because the low-rank model can help to remove the noise component and explore more of the essential structural information existing in the original clean data. Concurrently, our approach outperforms other low-rank-based feature learning methods and exhibits obvious improvement and robustness in the classification results when the data are corrupted by a heavy noise.

Discussion on parameters and convergence

There are several regularization parameters in our algorithm. In the following, we will briefly discuss them. Parameters μ and ρ are introduced owing to the ALM, and so they are set empirically as suggested in [28] to ensure convergence. For parameters η and λ in Eq (13), we choose COIL20 as the test dataset to study the effect on the classification results with their variational values. The classification curves for both the original data and their corrupted version versus η and λ are depicted in Figs 6 and 7, respectively. As can be seen from the results, the performance is insensitive to different η and λ, and it almost achieves consistent results over a wide range of these two parameters.

To verify the convergence of our approach, we plot the convergence curves of the objective function values versus the iterative steps in Fig 8. We choose Extended YaleB(Fig 8(A)) and COIL20(Fig 8(B)) as the testing datasets, and the settings are consistent with the experiments for the clean data. We can observe that our approach can well converge as the iteration proceeds.

Conclusion

In this paper, a robust and discriminative feature subspace learning method is proposed for feature extraction and classification tasks. Our approach iteratively learns a subspace with two types of constraints based on a low-rank representation and class labels, respectively. The ALM with BCD is developed to solve the framework convergently. The proposed approach is examined on several public datasets, and the experimental results demonstrate the competitive and superior performance of our approach compared to the conventional methods. In addition, when the data suffer from noise, our approach shows more robustness than the other comparison methods. In the future work, we may extend our approach to a semi-supervised scenario for feature learning and design some new regularization constraints to further improve the classification performance.

Supporting information

S1 Data. COIL20 test dataset.
(MAT)

Acknowledgments

The authors are grateful to the editor and anonymous reviewers for their valuable review comments on our work.

Author Contributions

Conceptualization: Ao Li, Xin Liu.

Formal analysis: Deyun Chen, Guanglu Sun.

Funding acquisition: Ao Li, Deyun Chen.

Investigation: Yanbing Wang, Kezheng Lin, Hailong Jiang.

Methodology: Ao Li, Xin Liu.

Validation: Ao Li, Xin Liu, Hailong Jiang.

Writing – original draft: Ao Li, Xin Liu.

Writing – review & editing: Ao Li, Yanbing Wang.

References

1. Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. null: IEEE, 2006: 2169–2178.
2. Bineng Z, Jun Z, Pengfei W, Jixiang D, Duansheng C. Jointly Feature Learning and Selection for Robust Tracking via a Gating Mechanism. PLOS ONE, 2016, 11(8): 1–15.
3. Li A, Wu ZQ, Lu HY, Chen DY, Sun GL. Collaborative self-regression method with nonlinear feature based on multi-task learning for image classification. IEEE Access, 2018, 6: 43513–43525.
4. Turk M, Pentland A. Eigenfaces for recognition. Journal of cognitive neuroscience, 1991, 3(1): 71–86. https://doi.org/10.1162/jocn.1991.3.1.71 PMID: 23964806
5. He X, Cai D, Yan S, Zhang HJ. Neighborhood preserving embedding. IEEE International Conference on Computer Vision, 2005: 1208–1213.
6. He X, Yan S, Hu Y, Niyogi P, Zhang HJ. Face recognition using laplacianfaces. IEEE transactions on pattern analysis and machine intelligence, 2005, 27(3):328–340. https://doi.org/10.1109/TPAMI.2005.55 PMID: 15747789
7. Lu YW, Yuan C, Li XL, Lai ZH, Zhang D, Shen LL. Structurally incoherent low-rank 2DLPP for image classification. IEEE Transactions on Circuits and Systems for Video Technology, 2018: 1–1.
8. Belhumeur PN, Hespanha JP, Kriegman DJ. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Yale University New Haven United States, 1997: 43–58.
9. Cai D, He X, Zhou K, Han J, Bao H. Locality Sensitive Discriminant Analysis. IJCAI, 2007: 1713–1726.
10. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(2): 210–227. https://doi.org/10.1109/TPAMI.2008.79 PMID: 19110489
11. Zhang L, Yang M, Feng X. Sparse representation or collaborative representation: Which helps face recognition? IEEE international conference on Computer vision, 2011: 471–478.
12. Li A, Chen DY, Wu ZQ, Sun GL, Lin KZ. Self-supervised sparse coding scheme for image classification based on low rank representation. PLOS ONE, 2018, 13(6): e0199141. https://doi.org/10.1371/journal.pone.0199141 PMID: 29924830
13. Zhang L, Zhu P, Hu Q, Zhang D. A linear subspace learning approach via sparse coding. IEEE international conference on Computer vision, 2011: 755–761.
14. Cai S, Zhang L, Zuo W, Feng X. A probabilistic collaborative representation based approach for pattern classification. IEEE conference on computer vision and pattern recognition, 2016: 2950–2959.
15. Li L, Li S, Fu Y. Learning low-rank and discriminative dictionary for image classification. Image and Vision Computing, 2014, 32(10): 814–823.
16. Zhou T, Liu F, Bhaskar H, Yang JJItoc. Robust visual tracking via online discriminative and low-rank dictionary learning, 2018, 48(9): 2643–2655.
17. Shao M, Kit D, Fu Y. Generalized transfer subspace learning through low-rank constraint. International Journal of Computer Vision, 2014, 109(1–2): 74–93.
18. Peng C, Kang Z, Cheng Q. Subspace clustering via variance regularized ridge regression. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 21–26.
19. Kang Z, Peng C, Cheng Q. Kernel-driven similarity learning. Neurocomputing, 2017, 267: 210–219.
20. Zhou X, Yang C, Yu W. Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(3): 597–610. https://doi.org/10.1109/TPAMI.2012.132 PMID: 22689075
21. Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y. Robust recovery of subspace structures by low-rank representation. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 171–184. https://doi.org/10.1109/TPAMI.2012.88 PMID: 22487984
22. Liu G, Yan S. Latent low-rank representation for subspace segmentation and feature extraction. IEEE international conference on Computer vision, 2011: 1615–1622.
23. Fang XZ, Han N, Wu JG, Xu Y, Yang J, Wong WK, et al. Approximate Low-Rank Projection Learning for Feature Extraction. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(11): 52288–5241.
24. Zhang Y, Jiang Z, Davis LS. Learning structured low-rank representations for image classification. IEEE conference on computer vision and pattern recognition, 2013: 676–683.
25. Ma L, Wang C, Xiao B, Zhou W. Sparse representation for face recognition based on discriminative low-rank dictionary learning. IEEE conference on computer vision and pattern recognition, 2012: 2586–2593.

26. Zhou P, Lin Z, Zhang C. Integrated low-rank-based discriminative feature learning for recognition. IEEE transactions on neural networks and learning systems, 2016, 27(5): 1080–1093. https://doi.org/10.1109/TNNLS.2015.2436951 PMID: 26080387
27. Luo T, Yang Y, Yi D, Ye J. Robust discriminative feature learning with calibrated data reconstruction and sparse low-rank model. Applied Intelligence, 2017: 1–14.
28. Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Preprint. Available from: arXiv: 1009.5055. Cited 2010.
29. Xu Y, Yin W. A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion. SIAM Journal on Imaging Science. 2013, 6(3): 1758–1789.
30. Liu G, Lin Z, Yu Y. Robust subspace segmentation by low-rank representation. Proceedings of the 27th international conference on machine learning, 2010: 663–670.
31. Kang Z, Wen LJ, Chen WY, Xu ZL. Low-rank kernel learning for graph-based clustering. Knowledge-Based Systems, 2019, 163: 510–517.
32. Peng C, Kang Z, Li HQ, Cheng Q. Subspace clustering using log-determinant rank approximation. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015: 925–934.
33. Wang YCF, Wei CP, Chen CF. Low-rank matrix recovery with structural incoherence for robust face recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2012: 2618–2625.
34. Gu Q, Wang ZR, Liu H. Low-rank and sparse structure pursuit via alternating minimization. Artificial Intelligence and Statistics, 2016: 600–609.
35. Keshavan RH, Montanari A, Oh S. Matrix completion from noisy entries. Journal of Machine Learning Research, 2010, 11(Jul): 2057–2078.
36. Cai JF, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 2010, 20(4): 1956–1982.
37. Zhuang L, Wang J, Lin Z, Yang AY, Ma Y, Yu N. Locality-preserving low-rank representation for graph construction from nonlinear manifolds. Neurocomputing, 2016, 175: 715–722.
38. Peng C, Kang Z, Cai S, Cheng Q. Integrate and Conquer: Double-Sided Two-Dimensional k-Means Via Integrating of Projection and Manifold Construction. ACM Transactions on Intelligent Systems and Technology. 2018, 9(5): 57.
39. Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. Journal of machine learning research, 2007, 8(May): 1027–1061.
40. Mika S, Ratsch G, Weston J, Scholkopf B, Mullers KR. Fisher discriminant analysis with kernels. Neural networks for signal processing IX: 1999 Proceedings of the 1999 IEEE signal processing society workshop, 1999: 41–48.
41. He X, Niyogi P. Locality preserving projections. Advances in neural information processing systems, 2004: 153–160.
42. Sugiyama M, Idé T, Nakajima S, Sese J. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. Machine learning, 2010, 78(1–2): 35.
43. Ioffe S. Probabilistic linear discriminant analysis. European Conference on Computer Vision: Springer, 2006: 531–542.
44. Clemmensen L, Hastie T, Witten D, Ersbøll B. Sparse discriminant analysis. Technometrics, 2011, 53 (4): 406–413.
45. Kwok JT, Wang TF, Liu TY. Large-scale low-rank matrix learning with nonconvex regularizers. IEEE transactions on pattern analysis and machine intelligence 2018: 1–1.
46. Wen Z, Yin W. A feasible method for optimization with orthogonality constraints. Mathematical Programming, 2013, 142(1–2): 397–434.
47. Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? Journal of the ACM (JACM), 2011, 58(3): 11.
48. Yang J, Zhang Y. Alternating direction algorithms for \ell_1-problems in compressive sensing. SIAM journal on scientific computing, 2011, 33(1): 250–278.
49. Yang J, Yin W, Zhang Y, Wang Y. A fast algorithm for edge-preserving variational multichannel image restoration. SIAM Journal on Imaging Sciences, 2009, 2(2): 569–592.