Kernel collaborative representation based dictionary learning and discriminative projection

Guoqing Zhang, Huaijiang Sun^{*}, [Guiyu Xia](#), Quansen Sun

School of Computer Science and Engineering, Nanjing University of Science and Technology
Jiangsu 210094, China

Abstract: Sparse representation classification (SRC) has been developed and shown great potential due to its effectiveness in various classification tasks. [But how to determine appropriate features that can best work with SRC remains an open question.](#) Based on SRC and [dimensionality reduction \(DR\)](#) techniques, [a simultaneous discriminative projection and dictionary learning method \(DSRC\) is proposed.](#) However, as a linear algorithm, DSRC cannot handle the data with highly nonlinear distribution. Recently research has shown that the collaborative representation mechanism is more important to the success of SRC. Motivated by [these concerns](#), in this paper, we propose a novel kernel collaborative representation based classifier (KCRC), and then we use it as a criterion to design a kernel collaborative representation based dictionary learning and discriminant projection method (KDL-DP). The proposed method aims at learning a projection matrix and a dictionary such that in the low dimension subspace the between-class reconstruction residual of a given data set is maximized and the within-class reconstruction residual is minimized. Extensive experimental results validate the superiority of the proposed approach when compared with the state-of-the-art methods.

Key word: Collaborative representation, dictionary learning, dimensionality reduction, kernel method.

1. Introduction

Dimensionality reduction (DR) is an important research topic in computer vision and pattern recognition fields. The high dimensional data usually lead to the inefficiency of many practical data processing techniques and may even degrade the performances of many classifiers. Many DR algorithms have been developed in the past decades [1-6]. The two most widely-used DR methods are principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2]. Both PCA and LDA are linear dimensionality reduction algorithms. They assume that the distributions of data sets are globally linear. However, high-dimensional data sets usually have non-linear structures in practical. Therefore, for conducting non-linear DR on a database, a family of manifold learning-related methods aroused wide research interests [3-6]. In addition, the kernel trick [7] has also been widely applied to extend linear DR algorithms to non-linear ones by performing linear operations on other higher or even infinite dimensional feature space induced by a kernel function. Yang et al. [8] proposed a complete kernel Fisher discriminant framework for feature extraction. Zafeiriou et al. [9] suggested a regularized kernel discriminant analysis method for face recognition and verification.

Recently, the idea of sparse representation is used to design some [DR](#) methods. Wright et al. [10] presented a sparse representation-based classification (SRC) method for face recognition. In SRC, a testing sample is sparsely reconstructed by the original training images and classified into the class with minimum sparse reconstruction residual. Qiao et al. [11] gave a sparsity preserving projections (SPP) for DR. In SPP, the graph (i.e. reconstruction relationships) was automatically constructed based on a modified sparse representation. Gui et al. [12] introduced the class information of data sets into

^{*} Corresponding author. [Tel:+86 13905172533](tel:+8613905172533)

Email address: sunhuaijiang@njust.edu.cn (Huaijiang Sun), xiayang14551@163.com (Guoqing Zhang)

SPP and presented a discriminant sparsity neighborhood preserving embedding (DSNPE). Yang et al. [13] claimed that DR algorithms should be designed according to the classifiers, so they gave a SRC steered discriminative projection (SRC-DP) method. By using SRC in the projected space, SRC-DP achieved better performance in face recognition. Clemmensen et al. [14] provided a sparse linear discriminant analysis (SLDA), which imposes sparseness constraint on projection vectors. The sparse projection vector yields a set of interpretable features for classification.

In sparse representation, the dictionary can be simply predefined as the training samples for discriminative sparse coding. However, it has been demonstrated that learning a dictionary from original training samples can provide better results [15-26]. These dictionary updating strategies are referred to as dictionary learning (DL). In [16], a DL algorithm, K-SVD, is introduced that generalizes k -means clustering and efficiently learns an overcomplete dictionary from a set of training samples. Since K-SVD algorithm focuses on the represent power without considering the discrimination power. Thus, based on K-SVD, Zhang et al. [17] developed a discriminative K-SVD algorithm by simultaneously learning a linear classifier. Jiang et al. [18, 19] introduced a label consistent regularization to enforce the discrimination of coding vectors. The so-called LC-KSVD algorithm exhibits good classification results. In order to further enhance the discrimination power of the learned dictionary, locality preserving K-SVD (LP-KSVD) [48] constructs a locality preserving term enforces the representation by local dictionary atoms similar to the input sample and penalizes the presentation by nonlocal (dissimilar) atoms. To exploit stronger discrimination, Yang et al. [21, 49] proposed a Fisher discrimination DL method, where the category-specific strategy is adopted for learning a structural dictionary and the Fisher discrimination criterion is imposed on the coding vectors to enhance class discrimination. In addition, Ramirez et al. [50] imposed a structure incoherence term on the dictionaries to make them independent to independent to each other as much as possible. Cai et al. [51] proposed and SVM guided DL algorithm, which weights the squared distances between all pairs of coding vectors, where the weights are determined adaptively by training an SVM. Yang et al. [52] proposed a latent DL method by jointly learning a latent matrix to adaptively build the relationship between dictionary atoms and class labels.

Recently, several jointly DR and DL methods [22, 53, 54] have been proposed and reported more competitive performance than conventional DL methods. Lu et al. [53] proposed a learning framework which simultaneously learns the feature and the dictionary for image set based face recognition. Feng et al. [54] jointly learns a DR matrix and a discriminative dictionary for face recognition. Zhang et al. [22] presented a simultaneous projection and DL method using a carefully designed sigmoid reconstruction error (the ratio of intra-class error to inter-class error in sigmoid function).

Most of sparsity based classification schemes need to solve the ℓ_1 -norm optimization problem, which is extremely time-consuming. Recent research in [27, 28, 45] have been shown that the collaborative representation (CR) mechanism is more important to the success of SRC. And a collaborative representation based classification (CRC) is given for image recognition by replacing the ℓ_1 -norm regularization with the ℓ_2 -norm regularization. CRC coded a testing sample by linear combination of all the training samples with regularized least square and classified the testing sample into the class with the minimum reconstruction error. Compared to SRC, CRC can

achieve the competitive classification performance with much less computation burden.

All the sparse representation-based or collaborative representation-based methods mentioned above are designed based on a linear representation of the data. It is always inadequate for representing non-linear structures of the data. To deal with this problem, some non-linear methods have been proposed in literature [29-33]. These methods map the non-linear data into high-dimensional feature space using the kernel trick so that data of the same type are easily grouped together and are linearly separable. In this case we may find the sparse representation for the data more easily, and the reconstruction error may be reduced significantly as well [29-33]. Kernel sparse representation-based classifier (KSRC) is a non-linear extension of SRC and can deal with the problem occurred in SRC [29]. Lai et al. [34] proposed a kernel locality-constrained collaborative representation based discriminant analysis (KLCRC-DA), which achieved better results on face recognition. The well-known K-SVD [16] and method of optimal directions (MOD) [46] learning algorithms have also been adapted to the high dimension feature space, and an efficient object recognition system that combines multiple classifiers based on the kernel sparse codes was presented in [33].

Motivated by the above concerns, in this paper, we propose a kernel collaborative representation based dictionary learning and discriminative projection method (KDL-DP). KDL-DP is designed according to the decision rule of our proposed kernel collaborative representation based classifier (KCRC), which is a nonlinear extension of CRC. The goal of the proposed method is to learn a projection matrix and a dictionary such that in the reduced subspace the within-class reconstruction residual is as small as possible and the between-class reconstruction residual is as large as possible. This encourages the sparse codes to be discriminative across classes. Therefore, KCRC can achieve better classification performance in the projected subspace.

The rest of paper is organized as follows: Section 2 reviews the CRC method. The kernel collaborative representation based classifier is proposed in Section 3. In Section 4, we design a kernel collaborative representation based dictionary learning and discriminative projection algorithm (KDL-DP). Experiments are described in Section 5 and Section 6 concludes the paper.

A note on notation: we use a bold upper case letter denotes a matrix, a bold lower case letter denotes a vector. For a matrix $\mathbf{P}$, $\mathbf{P}^T$ denotes its transpose. We summarize the key acronyms used throughout the paper in Table 1.

Table 1. The key acronyms of our paper

Notation	Meaning
DR	Dimensionality Reduction
DL	Dictionary Learning
SRC	Sparse Representation based Classification [10]
CRC	Collaborative representation based classification [27]
DSRC	Simultaneous discriminative projection and dictionary learning for sparse representation based classification [22]
SRC-DP	Sparse Representation Classifier Steered Discriminative Projection [13]
K-SVD	K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation [16]
LC-KSVD	Label Consistent K-SVD [19]
LP-KSVD	Joint kernel dictionary and classifier learning for sparse coding via locality

	preserving K-SVD [48]
FDDL	Fisher discrimination dictionary learning for sparse representation [21]
KSRC	Kernel Sparse Representation-based Classifier [29]
KCRC	Kernel collaborative representation based classifier
KDL-DP	Kernel collaborative representation based dictionary learning and discriminative projection

2. Collaborative representation based classifier

In this section, we briefly review collaborative representation based classification (CRC). Suppose we have c classes of subjects. Denote $\mathbf{X}_i = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_i}] \in R^{m \times n_i}$ as the training dataset of the i -th class, where $\mathbf{x}_j \in R^m, j=1, \dots, n_i$, is a m -dimensional vector stretched by the j -th sample of class i . The entire training set is defined as $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_c] \in R^{m \times n}$, where $n = \sum_{i=1}^c n_i$. Denote $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_c] \in R^{m \times q}$ as the **structured** dictionary, where $\mathbf{D}_i \in R^{m \times q_i}$ is the sub-dictionary associated with the i -th class $\mathbf{X}_i$, and $q = q_1 + \dots + q_c$. Once a test sample $\mathbf{y}$ comes, we **can** represent $\mathbf{y}$ by using the dictionary, namely $\mathbf{y} = \mathbf{D}\mathbf{a}$, where $\mathbf{a} = [\mathbf{a}_1; \mathbf{a}_2; \dots; \mathbf{a}_c]$ and $\mathbf{a}_i$ is the reconstruction representation vector associated with class i . If $\mathbf{y}$ from the i -th class, usually $\mathbf{y} = \mathbf{D}_i\mathbf{a}_i$ holds well. This implies that most coefficients in $\mathbf{a}$ are nearly zeros and only $\mathbf{a}_i$ has significant nonzero entries. Then the coding coefficient $\mathbf{a}$ can be obtained by solving the following optimization problem:

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\mathbf{a}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{D}\mathbf{a}, \quad (1)$$

where $\|\cdot\|_0$ is the ℓ_0 -norm, which counts the number of nonzero entries in a vector.

Solving ℓ_0 optimization problem in Eq. (1), however, is NP hard and extremely time-consuming. Fortunately, recent research efforts reveal that for certain dictionaries, if the solution is sparse enough, finding the solution of the ℓ_0 optimization problem is equivalent to a ℓ_1 optimization problem [35, 36].

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\mathbf{a}\|_1 \quad s.t. \quad \mathbf{y} = \mathbf{D}\mathbf{a}. \quad (2)$$

In practice, Eq. (2) is always transferred to a regularization problem, which is also called Lasso [37].

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1, \quad (3)$$

where λ is the regularization parameter. Solving the ℓ_1 -minimization problem in Eq. (3) is still time

consuming. Recently, Zhang et al pointed out that the use of collaborative representation was more crucial than the ℓ_1 -sparsity of $\mathbf{a}$ [27, 45]. So they proposed the collaborative representation based classification (CRC) method which tried to solve the following problem:

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_2^2, \quad (4)$$

where $\|\cdot\|_2$ denotes the ℓ_2 -norm, and $\hat{\mathbf{a}} = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$ is considered as the collaborative representation of $\mathbf{y}$ in terms of $\mathbf{D}$.

The solution of CR with regularized least square in Eq. (4) can be easily and analytically derived as $\hat{\mathbf{a}} = (\mathbf{D}^T \mathbf{D} + \lambda \mathbf{I})^{-1} \mathbf{D}^T \mathbf{y}$.

The classification scheme is similar to SRC [2]. Once $\hat{\mathbf{a}}$ is obtained, the regularized representation error for each class is computed by $r_i = \|\mathbf{y} - \mathbf{D}_i \hat{\mathbf{a}}_i\|_2 / \|\hat{\mathbf{a}}_i\|_2$, where $\hat{\mathbf{a}}_i$ is the coding coefficient associated with class i . The classification is made by $\text{identify}(\mathbf{y}) = \arg \min_i (r_i)$.

3. Kernel collaborative representation-based classifier

Let $\phi: \mathbf{x} \rightarrow \phi(\mathbf{x})$ be a non-linear mapping from original feature space R^m into a high dimensional feature space H_o . Then the training sample matrix, the testing sample $\mathbf{y}$ and the representation dictionary $\mathbf{D}$ in space H_o can be expressed as $\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \dots, \phi(\mathbf{x}_n)]$, $\phi(\mathbf{y})$ and $\tilde{\mathbf{D}} = [\phi(\mathbf{d}_1), \phi(\mathbf{d}_2), \dots, \phi(\mathbf{d}_q)]$, respectively. Eq. (4) can be changed into the following formula:

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\phi(\mathbf{y}) - \tilde{\mathbf{D}}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_2^2 \quad (5)$$

Since the feature space H_o has a very high or possibly infinite dimensionality, it is necessary to perform dimensionality reduction in H_o [29, 34]. In other words, the implicit samples need to be projected from H_o into a low-dimensional (reduced) subspace. Suppose we have the projection matrix

$\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_d] \in R^{\tilde{m} \times d}$ (d is the dimension of projection space), then Eq. (5) can be reformulated as

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\mathbf{P}^T \phi(\mathbf{y}) - \mathbf{P}^T \tilde{\mathbf{D}}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_2^2 \quad (6)$$

For the high dimensional feature space induced, $\mathbf{P}$ could be expressed $\mathbf{P} = \phi(\mathbf{X})\mathbf{B}$, where $\phi(\mathbf{X})$ is the data matrix in the feature space and $\mathbf{B} \in R^{n \times d}$ is the coefficient matrix for $\mathbf{P}$. Since dictionary atoms lie within the subspace spanned by the transformed samples $\phi(\mathbf{X})$, we can represent the dictionary

$\tilde{\mathbf{D}} = \phi(\mathbf{X})\mathbf{A}$. Through some derivations, Eq. (6) can be expressed as follows:

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \|\mathbf{B}^T \mathbf{K}(\cdot, \mathbf{y}) - \mathbf{B}^T \mathbf{K} \mathbf{A} \mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_2^2 \quad (7)$$

where $\mathbf{K}$ is the kernel Gram matrix of $\phi(\mathbf{X})$ which is symmetric and positive semi-definite,

$\mathbf{K}(\cdot, \mathbf{y}) = [k(\mathbf{x}_1, \mathbf{y}), \dots, k(\mathbf{x}_n, \mathbf{y})]^T$, and $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$ is a special Mercer kernel. Solving Eq. (7), we

can get the coefficient vector $\hat{\mathbf{a}} = (\mathbf{A}^T \mathbf{K}^T \mathbf{B} \mathbf{B}^T \mathbf{K} \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{K}^T \mathbf{B} \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{y})$. Now we need to classify $\mathbf{y}$

in terms of $\hat{\mathbf{a}}$. We use the minimum residual between $\mathbf{y}$ and its c approximations in the reduced subspace to determine the label of $\mathbf{y}$. The residuals of different classes can be computed as

$$r_i(\mathbf{y}) = \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{y}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_i \hat{\mathbf{a}}_i \right\|_2 / \left\| \hat{\mathbf{a}}_i \right\|_2, \text{ for } i=1, 2, \dots, c, \quad (8)$$

where $\mathbf{A}_i$ is the i -th sub-matrix of $\mathbf{A}$, and $\mathbf{a}_i$ is the coefficient vector associated with class i ($i=1, 2, \dots, c$). Then $\mathbf{y}$ is assigned to the corresponding class with minimal residual.

4. Kernel collaborative representation based dictionary learning and discriminative projection

4.1 Formulation

In this section, a kernel collaborative representation based dictionary learning and discriminative projection (KDL-DP) method is proposed. The proposed method is designed according to the decision rule of KCRC and the basic idea of the proposed method is to seek a projection matrix and a dictionary such that the reconstruction residual for the within-class should be small while all others should be large in the reduced subspace. Hence, we can learn discriminative projection matrix and dictionary simultaneously, and KCRC can achieve satisfactory performance in the low dimensional subspace. In addition, we will show that KDL-DP and KCRC depend on each other. It means that the recognition coefficient vectors required in KDL-DP need to be computed by KCRC and the projection matrix and dictionary required in KCRC need to be computed by KDL-DP. Therefore, we use an alternative iterative approach to obtain the optimal solution of the proposed algorithm.

For each training sample $\phi(\mathbf{x})$ which belongs to class i , we define the within-class reconstruction residual in the projected subspace as $e_i = \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_i \mathbf{a}_i \right\|_2^2$. And we define the between-class reconstruction residual as the smallest reconstruction residual from classes other than i (denoted as class j): $e_j = \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_j \mathbf{a}_j \right\|_2^2$, where j can be obtained as

$$j = \arg \min_s e_s = \arg \min_s \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_s \mathbf{a}_s \right\|_2^2 \text{ s.t. } s \neq i \quad (9)$$

Given the training set $\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \dots, \phi(\mathbf{x}_n)]$, the task is to learn a discriminative projection $\mathbf{P}$ via $\mathbf{B}$ and a dictionary $\tilde{\mathbf{D}}$ via $\mathbf{A}$ in the reduced subspace. The objective function can be formulated as follows:

$$J(\mathbf{A}, \mathbf{B}) = \arg \min_{\mathbf{A}, \mathbf{B}} \frac{1}{n} \sum_{l=1}^n \left(S_\theta(R(\mathbf{x}_l)) + \lambda \left\| \mathbf{a}_l \right\|_2^2 \right) \quad (10)$$

where $\mathbf{a}_l$ is the representation coefficient of $\mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}_l)$ with respect to $\mathbf{B}^T \mathbf{K} \mathbf{A}$,

$$\hat{\mathbf{a}}_l = \arg \min_{\mathbf{a}} \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}_l) - \mathbf{B}^T \mathbf{K} \mathbf{A} \mathbf{a} \right\|_2^2 + \lambda \left\| \mathbf{a} \right\|_2^2 \quad (11)$$

$S_\theta(x)$ is a sigmoid function centered at 1 with slope θ which is used as a soft approximation to the step

function (see Figure 1) [22] and is defined as $S_\theta(x) = \frac{1}{1 + e^{\theta(1-x)}}$, $R(\mathbf{x}_l)$ is defined as

$$R(\mathbf{x}_l) = \frac{e_l^i}{e_l^j} \quad (12)$$

which actually reflects the relative difference between the within-class residual and between-class residual. As described in [22], when $R(\mathbf{x}_i) < 1$, the test sample $\mathbf{x}_i$ is classified correctly, which means the within-class gives the smallest reconstruction error. If $R(\mathbf{x}_i) > 1$, then the reconstruction residual for the within-class is larger than that of class j , therefore x_i is miss-classified as class j [22]. Intuitively, $S_\theta(R(\mathbf{x}_i))$ can be regarded as inverse classification confidence [22], thus, to learn the discriminative projection and dictionary for improving classification, we should minimize this inverse confidence empirically over the training data. By minimizing the objective function in Eq. (10), we can learn the discriminative projection $\mathbf{P}$ and dictionary $\tilde{\mathbf{D}}$ jointly.

To minimize the objective function (10), we use an iterative approach to obtain the optimal solution of $\mathbf{B}$ and $\mathbf{A}$. The optimization scheme is based on gradient descent.

Fig. 1. Illustration of the sigmoid function $S_\theta(\mathbf{x})$ with different values for the slope parameter θ

4.2 Optimization method

Our optimization algorithm alternates between sparse coding and dictionary/projection updating. It means that the reconstruction coefficient vectors required in KDL-DP need to be computed by KCRC and the projection matrix and dictionary required in KCRC need to be computed by KDL-DP. Given a training set $\{\mathbf{x}_i\}_{i=1}^n$, we compute the sparse coding for each training sample $\mathbf{x}_i$ with the current estimation of the $\mathbf{B}$ and $\mathbf{A}$ by solving Eq. (11). Then, the new projection/dictionary is updated via minimizing Eq. (10) given sparse codes $\{\mathbf{a}_i\}_{i=1}^n$. From Eq. (10), we can get the following expressions for gradients with respect to $\mathbf{B}$ and $\mathbf{A}$ (denoted as $g_{\mathbf{B}}$ and $g_{\mathbf{A}}$ respectively) [22]:

$$\begin{aligned}
 g_{\mathbf{B}} &= \frac{\partial J}{\partial \mathbf{B}} = \sum_i \frac{\partial S_\theta(R(\mathbf{x}_i))}{\partial \mathbf{B}} = \sum_i \frac{\partial S_\theta(R(\mathbf{x}_i))}{\partial R(\mathbf{x}_i)} \frac{\partial R(\mathbf{x}_i)}{\partial \mathbf{B}} \\
 &= \sum_i \frac{\partial S_\theta(R(\mathbf{x}_i))}{\partial R(\mathbf{x}_i)} \left(\frac{\partial R(\mathbf{x}_i)}{\partial e_i^i} \frac{\partial e_i^i}{\partial \mathbf{B}} + \frac{\partial R(\mathbf{x}_i)}{\partial e_i^j} \frac{\partial e_i^j}{\partial \mathbf{B}} \right) \\
 &= \sum_i \frac{S_\theta'(R(\mathbf{x}_i))}{e_i^j} \nabla_{\mathbf{B}} e_i^i - \sum_i \frac{S_\theta'(R(\mathbf{x}_i)) e_i^i}{(e_i^j)^2} \nabla_{\mathbf{B}} e_i^j
 \end{aligned} \tag{13}$$

$$\begin{aligned}
g_{\mathbf{A}} &= \frac{\partial J}{\partial \mathbf{A}} = \sum_l \frac{\partial S_{\theta}(R(\mathbf{x}_l))}{\partial \mathbf{A}} = \sum_l \frac{\partial S_{\theta}(R(\mathbf{x}_l))}{\partial R(\mathbf{x}_l)} \frac{\partial R(\mathbf{x}_l)}{\partial \mathbf{A}} \\
&= \sum_l \frac{\partial S_{\theta}(R(\mathbf{x}_l))}{\partial R(\mathbf{x}_l)} \left(\frac{\partial R(\mathbf{x}_l)}{\partial e_l^i} \frac{\partial e_l^i}{\partial \mathbf{A}} + \frac{\partial R(\mathbf{x}_l)}{\partial e_l^j} \frac{\partial e_l^j}{\partial \mathbf{A}} \right) \\
&= \sum_l \frac{S_{\theta}'(R(\mathbf{x}_l))}{e_l^j} \nabla_{\mathbf{A}} e_l^i - \sum_l \frac{S_{\theta}'(R(\mathbf{x}_l)) e_l^i}{(e_l^j)^2} \nabla_{\mathbf{A}} e_l^j
\end{aligned} \tag{14}$$

Recall that e^g ($g \in \{i, j\}$) represents the reconstruction error in the reduced subspace and is defined as

$$e^g = \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_g \hat{\mathbf{a}}_g \right\|_2^2 \tag{15}$$

Therefore

$$\nabla_{\mathbf{B}} e^g = 2(\mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_g \hat{\mathbf{a}}_g)(\mathbf{K}(\cdot, \mathbf{x}) - \mathbf{K} \mathbf{A}_g \hat{\mathbf{a}}_g)^T \tag{16}$$

$$\nabla_{\mathbf{A}} e^g = -2\mathbf{K}_g \mathbf{B}(\mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_g \hat{\mathbf{a}}_g) \hat{\mathbf{a}}_g^T \tag{17}$$

Note that as $\mathbf{a}$ is related to $\mathbf{B}$ and $\mathbf{A}$ through Eq. (11), there should be a term that reflects the contribution from $\mathbf{a}$ to $\nabla_{\mathbf{B}} e^g$ [38 - 40]. In practice, we use an alternating minimization scheme and treats $\mathbf{a}$ as a known variable when calculating the gradient of e^g with respect to $\mathbf{B}$ and $\mathbf{A}$, thus inducing the equation as shown in Eq. (17). The derivative of $S_{\theta}(x)$ can be calculated as follows:

$$S_{\theta}(x) = \frac{dS_{\theta}(x)}{dx} = \frac{\theta e^{\theta(1-x)}}{(1 + e^{\theta(1-x)})^2} \tag{18}$$

Putting Eq. (16)-Eq. (18) into Eq. (13) and Eq. (14), we can get the full expression of the gradient direction for updating $\mathbf{B}$ and $\mathbf{A}$. The overall optimization procedure is summarized in Algorithm 1. For the given training set $\{\mathbf{x}_l\}_{l=1}^n$, we first compute the sparse representation coefficients with respect to the current $\mathbf{B}$ and $\mathbf{A}$. Then we calculate the gradient descent directions for updating the projection and dictionary.

Algorithm 1 KDL-DP

Input: A set of training samples $\{\mathbf{x}_l\}_{l=1}^n$, the label of $\mathbf{x}_l$ is $l_i \in \{1, \dots, c\}$, A kernel function k .

Initialize: we use KPCA to initialize $\mathbf{P}$ and get the corresponding pseudo-transformation matrix $\mathbf{B}^{(0)}$.

Set a random element of each column in $\mathbf{A}^{(0)}$ to be 1. Normalize each column of $\mathbf{B}^T \mathbf{K}$ and $\mathbf{B}^T \mathbf{K}(\cdot, \mathbf{x})$ to have unit ℓ_2 -norm. Set iteration $t = 1$.

Repeat

for $l = 1, \dots, n$ **do**

 Compute coding coefficient via (11) to get $\mathbf{a}_l$.

 Get j_l via (9).

end

 Compute $g_{\mathbf{B}}$ for $\mathbf{B}$ via (13), $g_{\mathbf{A}}$ for $\mathbf{A}$ via (14).

$$\text{Update } \mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} - \gamma g_{\mathbf{B}}, \mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} - \eta g_{\mathbf{A}}.$$

$t = t + 1$.

Until convergence

Output: $\mathbf{B}$ and $\mathbf{A}$

4.3 Classification Strategy

Once we learned the projection and kernel dictionary, we need to project the query image via learned projection, code the projected sample over the learned kernel dictionary and eventually obtain its coding coefficient which is expected to be discriminative. We first project the query sample and then code it over the learning kernel dictionary via

$$\begin{aligned} \hat{\mathbf{a}} &= \arg \min_{\mathbf{a}} \left\| \mathbf{P}^T \phi(\mathbf{y}) - \mathbf{P}^T \tilde{\mathbf{D}} \mathbf{a} \right\| + \lambda \left\| \mathbf{a} \right\|_2^2 \\ &= \arg \min_{\mathbf{a}} \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{y}) - \mathbf{B}^T \mathbf{K} \mathbf{A} \mathbf{a} \right\|_2^2 + \lambda \left\| \mathbf{a} \right\|_2^2 \end{aligned} \quad (19)$$

where $\hat{\mathbf{a}} = [\hat{\mathbf{a}}_1; \hat{\mathbf{a}}_2; \dots; \hat{\mathbf{a}}_c]$ and $\hat{\mathbf{a}}_i$ is the coefficient vector associated with class i . After obtaining the coding coefficients $\hat{\mathbf{a}}$, we need to classify $\mathbf{y}$ in terms of $\hat{\mathbf{a}}$. The residuals of different classes can be computed as

$$r_i(\mathbf{y}) = \left\| \mathbf{B}^T \mathbf{K}(\cdot, \mathbf{y}) - \mathbf{B}^T \mathbf{K} \mathbf{A}_i \hat{\mathbf{a}}_i \right\|_2 / \left\| \hat{\mathbf{a}}_i \right\|_2 \quad (20)$$

The final classification is performed by

$$\text{label}(\mathbf{y}) = \arg \min_i r_i(\mathbf{y}) \quad (21)$$

4.4 Convergence of the KDL-DP method

The proposed learning algorithm is a gradient based approach. While there is no guarantee of obtaining the global optimal solution. Empirically, satisfactory solutions to the desired kernel dictionary $\tilde{\mathbf{D}}$ and the projection matrix $\mathbf{P}$ can be obtained with the decreasing of the objective function. Let us use the AR face database as an example to illustrate the optimization process of KDL-DP (the experiment setting is given in next section), the dimensionality of samples is reduced to 200, the number of atoms for each class is 7. The curve of the objective function vs. the number of iterations is plotted in Fig. 2. From Fig. 2, we can see that after several iterations (e.g., 25 iterations), the value of the objective function becomes stable, and it varies only in a small range. Therefore, the iteration number in our experiments usually set as less than 30.

Fig. 2. An example of convergence of KDL-DP on AR face database.

5. Experiment and analysis

In this section, the performance of the proposed KDL-DP method is evaluated by a series of experiments on face, object, handwritten and digits datasets. In order to clearly illustrate the advantage of the proposed method, the related algorithms including CRC [27], SRC-DP [13], KSRC [29], LC-KSVD [19], Fisher discrimination dictionary learning (FDDL) [21], DSRC [22] and LP-KSVD [48] are used for comparison. The recognition rate is used as the criterion to compare the performance of different state-of-the-art algorithms.

It should be noted that, SRC-DP is devised based on SRC. In our experiments, we hope to test the feature extraction abilities of different subspace learning algorithms for CRC, so the ℓ_1 -minimization problems solved in SRC-DP will be replaced with ℓ_2 -minimization in our experiments (denoted as CRC-DP). In addition, for CRC, CRC-DP and KSRC, the number of atoms for each class is set to the same as the number of training samples of each class.

5.1 Implementation details

In this subsection, we provide implementation details of our approach. There are five parameters in Algorithm 1: $t, \lambda, \theta, \gamma$ and η . We know that parameter determination is a big challenge in pattern recognition and computer vision. How to set the parameters' values is an important task, to examine the influence of parameters over the recognition results, we carried out some experiments using different databases. t is the number of iterations, according to the convergence of our algorithm (subsection 4.4), we empirically observe that the algorithm typically converge in less than 30 iterations, therefore, in our experiments we set $t = 30$. The sparsity regularization parameter λ are fixed for each data set and determined by 5-fold cross validation on the training data. The learned values for different datasets are presented in the following experiment section. θ controls the slope of sigmoid function $S_\theta(x)$ (see Fig. 1), we adjust the parameter θ within the range of $\theta \in [2, 10]$. γ is the step size for updating the projection, which is set as $\gamma \in [0.1, 0.5]$. η is the step size for updating the dictionary, we fix it as $\eta = 0.05$ in all the experiments. Fig. 3 illustrates the recognition rates of KDL-DP with varying θ , γ and η on different datasets (the other experiment settings are given in next subsections). We vary the value of one parameter and fixed the value of the others. From the results, we can see that we chose the appropriate parameter values for θ , γ and η . The Gaussian

kernel $\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{c}\right)$ is used in our method, with $c = \frac{1}{\left(\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2\right)}, i=1, \dots, n$, where $\bar{\mathbf{x}}$ is the mean

of all training samples. Each experiment was repeated 10 times to calculate the average recognition rate.

Fig. 3. Recognition rates of KDL-DP with varying θ , γ and η on different datasets. (a) The recognition rates versus the variation of the slope of sigmoid function $S_\theta(x)$. (b) The recognition rates versus the variation of the step size for updating the projection. (c) The recognition rates versus the variation of the step size for updating the dictionary.

5.2 Face recognition

In this subsection, we use the public AR, Extended Yale B and CMU PIE databases to validate the effectiveness of the proposed approach.

5.2.1 Experiment Using the AR database

The AR face database [41] contains over 4000 color face images of 126 people, including frontal views of faces with different expressions, lighting conditions, and occlusions. The pictures of 120 individuals (65 men and 55 women) are taken in two sessions (separated by two weeks) and each session contains 13 color images. Fourteen face images without occlusion (each session containing seven) of these 120 individuals are selected and used in our experiment. The face portion of each image is manually cropped and then normalized to 50×45 pixels. The sample images of one person are shown in Fig 4.

Fig. 4. Some samples from the AR database

In our experiments, we randomly select half of images in each subject as the training samples, and the remaining as the testing samples, to avoid over-fitting, we firstly perform PCA and reduce the dimension to be 200 before implementing the evaluating algorithms and vary the number of atoms (dictionary size) for each sub-dictionary q_i within the set $q_i \in \{2, 4, 6, 7\}$. The experimental results are summarized in Table 2. As can be seen from Table 2, the proposed method is very effective under different dictionary sizes. When the dictionary size is very small $q_i = 2$, the proposed method performs

well, and outperforms other algorithms. With the increase of the number of atoms, the performances of all the algorithms also increase in general and our method still performs the best. λ are determined by fivefold cross validation on the training dataset. We can observe that good performance is achieved at $\lambda = 0.001, \gamma = 0.5, \theta = 10$.

We then evaluate the performance of all the algorithms under different dimensionalities of features. The experiment results are shown in Fig 5. From Fig 5, we can see that the proposed method consistently outperforms the related algorithms. Especially when the dimensionality of the features is low, the performance of KDL-DP is much better than those of the other algorithms, which implying its effectiveness in capturing discriminative information **with low feature dimension** for classification. So that a more effective classifier can be obtained.

Table 2: Recognition rate (%) using different number of atoms on the AR database

Method	2 atoms	4 atoms	6 atoms	7 atoms
CRC [27]	71.39	85.97	91.62	92.37
CRC-DP [13]	75.30	84.03	89.52	91.46
KSRC [29]	74.83	88.88	93.02	93.76
LC-KSVD [19]	76.34	86.82	90.98	91.91
FDDL [21]	75.68	88.76	92.19	93.27
DSRC [22]	77.55	89.07	93.29	94.38
LP-KSVD (kernel) [48]	76.99	87.67	92.69	94.05
KDL-DP	80.65	90.91	95.04	95.93

Fig. 5. Recognition rates (%) under different dimensionality of features on the AR database. The number of atoms per class is (a) 2, (b) 4, (c) 6, (d) 7, respectively.

5.2.2 Experiment Using the Extended Yale B database

The Extended Yale B database contains 2,414 frontal face images of 38 people. There are about

64 images for each person. The cropped and normalized 192×168 face images were captured under various laboratory-controlled lighting conditions (see Fig. 6 for example). For computational convenience, these images were resized to 48×42 .

Fig. 6. Some samples from the Extended Yale B database.

In this experiment, we randomly select half of images in each subject as the training samples, and the remaining as the testing samples. As we do in the previous experiments, PCA is first performed and the dimension of original data is reduced to be 200 before implementing the evaluating algorithms. We vary the number of atoms from 8 to 32 with increments of 4. Moreover, in order to robustly evaluate the performances of different algorithms, we select images randomly to calculate the average recognition rate. $\lambda = 0.05, \gamma = 0.5, \theta = 10$ are used in this experiment. The experiment results are shown in Fig 7. We can draw similar conclusions to those on the AR database, i.e., the proposed KDL-DP method is more remarkable when the number of atoms is low. In general, the performance difference between KDL-DP and others become less and less significant with the increase of dictionary atoms.

We then evaluate the performance of KDL-DP on different dimensionalities of features. We select the number of atoms $q = 20$. The recognition rates with different dimensions are shown in Fig 8. It can be seen that the proposed method consistently outperforms the related algorithms, especially when the dimensionality of the features is low.

Fig. 7. Average recognition rates (%) under different number of dictionary atoms on Yale B database

Fig. 8. Recognition rates under different dimensionality of features (20 atoms per class)

5.2.3 Experiment Using the PIE database

The CMU PIE face database [44] contains 68 subjects with over 40 000 face images. Images of each person are taken across 13 different poses, under 43 different illumination conditions, and with four different expressions. Here we use a subset containing images of pose C05 (a nearly frontal pose of 68 persons, each with 49 images). All images are manually aligned, cropped, and resized to be 64×64 pixels in our experiments. Some images of one person are shown in Fig. 9.

Fig. 9. Some samples of one person from the CMU PIE dataset

By using the same methodology as adopted in section 5.2.2, we randomly select half of images in each subject as the training samples, and the remaining as the test samples. PCA is first performed and the dimension of original data is reduced to be 200 before implementing the evaluating algorithms. We vary the number of atoms from 5 to 20 with increments of 5. We repeated this setting 10 times and the average recognition rates are shown in Fig. 10. From Fig. 10 we can draw similar conclusions to those on the AR and Extended Yale B databases. i.e., the proposed KDL-DP achieves the best recognition rates and its advantage over the other methods is more remarkable when the number of training samples is less sufficient.

Fig. 10. Average recognition rates (%) under different number of dictionary atoms on PIE database

5.3 Object recognition

The COIL-20 database [42] contains 1440 grayscale images of 20 objects (72 images per object) under various poses. These objects have a wide variety of complex geometric, appearance and reflectance characteristics. They are rotated through 360 degrees against a black and taken at the interval of 5° . Each image is down-sampled to the size of 32×32 . Fig. 11 shows twenty different object images from the COIL-20 database.

In this experiment, we randomly select half of images in each subject as the training samples, and the remaining as the test samples. The number of atoms for each class in the dictionary is $q = 20$. The classification results under different numbers of dimensions are reported in Table 3. As can be seen from Table 3, the proposed method outperforms [linear dictionary learning methods](#): FDDL, DSRC and [LC-KSVD](#) as well as [kernel-based methods](#): KSRC, and LP-KSVD, irrespective of the variations of dimensions. When the number of dimensions is relatively low, KDL-DP significantly outperforms other methods. This indicates that simultaneously learned projection can preserve discriminative information with low feature dimension, leading to competitive performance of KDL-DP.

Fig. 11. Twenty object samples from the COIL-20 dataset. For each object, only one image is shown

Table 3 Recognition rate (%) under different dimensionalities of features with dictionary size $q = 20$

Dim	CRC	CRC-DP	KSRC	LC-KSVD	FDDL	DSRC	LP-KSVD (kernel)	KDL-DP
30	78.50	80.62	81.26	79.95	73.12	83.47	81.63	86.50
50	84.12	85.53	88.49	84.29	83.49	90.18	88.25	92.29
100	91.33	92.48	92.6	90.26	89.76	94.34	93.00	95.13
200	94.35	94.06	95.14	93.41	94.75	95.53	95.37	96.38

5.4 Digit Recognition

In this experiment, we focus on the digit recognition task using the USPS [43] handwritten digits

database. The USPS database consists of 7291 training images and 2007 test images of 10 classes. The image of size is 16×16 . For USPS database, the recognition results with dictionary size of $q = 100$ under different numbers of dimensions are reported in Table 3. λ is set to 0.001 in the experiment and $\gamma = 0.25, \theta = 10$. From Table 4 we can observe that KDL-DP achieves the highest recognition rates in most cases. FDDL obtain the highest performance when the feature dimension is 100, and our method obtains the second best performance. This indicates that in most cases our non-linear dictionary learning method outperforms the linear and the non-linear dictionary learning methods and also outperforms other non-dictionary learning methods.

Table 4 Recognition rate (%) under different dimensionalities of features with dictionary size $q = 100$

Dim	CRC	CRC-DP	KSRC	LC-KSVD	FDDL	DSRC	LP-KSVD (kernel)	KDL-DP
30	86.49	87.39	86.62	85.59	83.21	88.33	88.46	91.63
50	90.65	90.57	91.51	92.28	91.24	92.02	92.96	93.37
100	93.54	94.98	95.68	94.60	96.07	93.02	95.21	95.25
200	92.41	94.43	96.14	94.00	96.31	95.27	95.88	96.51

6. Conclusion

In this paper, we present a novel kernel collaborative representation based classifier (KCRC). Based on the classification rule of KCRC, a kernel collaborative representation based dictionary learning and discriminative projection method is proposed. We used a gradient descent based algorithm to learn a discriminative projection and a dictionary by maximizing the difference between the within-class and between-class reconstruction residual such that in the reduced subspace KCRC achieved better performance and become more efficient. Experiments have been conducted on various databases with the proposed method. The experiment results show that KDL-DP surpasses many state-of-the arts methods.

Acknowledgments

This work was supported by the National Science Foundation of China under Grants 61273251 and 61401209, in part by Project of Civil Space Technology Preresearch of the 12th Five-Year Plan, in part by the Natural Science Foundation of Jiangsu Province, China under Grant BK20140790, and in part by China Postdoctoral Science Foundation under Grants 2014T70525 & 2013M531364.

References

- [1] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1) (1991) 71-86.
- [2] P.N. Belhumeur, J. P. Hespanha, D.J. Kriegman, Eigenface vs. Fisherfaces: Recognition Using Class specific Linear Projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (7) (1997) 711-720.
- [3] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face Recognition Using Laplacianfaces, *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3) (2005) 328-340.
- [4] J. Tenenbaum, V. Silva, J. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, 290(22) (2000) 2319-2323.
- [5] S. Roweis, L. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, 290(22) (2000) 2323-2326.

- [6] M. Belkin, P. Niyogi, Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, *Advances in Neural Information Processing System*, 14 (2001) 585-591.
- [7] K. Muller, S. Mika, G. Riitsch, K. Tsuda, B. SchÖlkopf, An Introduction to Kernel-Based Learning Algorithms, *IEEE Trans. Neural Netw. Learn. Syst.*, 12 (2001) 181-201.
- [8] J. Yang, A.F. Frangi, J. Y. Yang, D. Zhang, Z. Jin, KPCA plus LDA: A complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2) (2005) 230-244.
- [9] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, T. Stathaki, Regularized Kernel Discriminant Analysis with A Robust Kernel for Face Recognition and Verification, *IEEE Trans. Neural Netw. Learn. Syst.*, 23 (3) (2012) 526-534.
- [10] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust Face Recognition via Sparse Representation, *IEEE Trans. Pattern Anal. Mach. Intell.*, 31 (2) (2009) 210-227.
- [11] Qiao L, Chen S, Tan X, Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 43(1) (2010) 331-341.
- [12] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, S. Ji, Discriminant Sparse Neighborhood Preserving Embedding for Face Recognition, *Pattern Recognition*, 45 (2012) 2884-2893.
- [13] J. Yang, D. Chu, L. Zhang, Y. Xu, J. Yang, Sparse Representation Classifier Steered Discriminative Projection with Applications to Face Recognition, *IEEE Trans. Neural Netw. Learn. Syst.*, 24(7) (2013) 1023-1035.
- [14] L. Clemmensen, T. Hastie, D. Witten, B. Ersboll, Sparse Discriminant Analysis, *Technometrics*, 53 (4) (2011) 406-413.
- [15] J. Mairal, F. Bach, J. Ponce, Task-Driven Dictionary Learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, 34 (11) (2012) 791-804.
- [16] M. Aharon, M. Elad, A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. Signal Processing* 54(1) (2006) 4311-4322.
- [17] Q. Zhang, B. Li, Discriminative K-SVD for dictionary learning in face recognition, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2691-2698.
- [18] Z. Jiang, Z. Lin, L. Davis, Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1697-1704.
- [19] Z. Jiang, L. Davis, Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11) (2013) 2651-2664.
- [20] M. Yang, L. Zhang, J. Yang, D. Zhang, Metaface Learning for Sparse Representation based Face Recognition, in: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 1601-1604.
- [21] M. Yang, L. Zhang, X. Feng, D. Zhang, Fisher discrimination dictionary learning for sparse representation, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 543-550.
- [22] H. Zhang, Y. Zhang, T.S. Huang, Simultaneous discriminative projection and dictionary learning for sparse representation based classification, *Pattern Recognition*, 46 (1) (2013) 346-354.
- [23] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Supervised dictionary learning, in: *Advances in Neural Information Processing Systems (NIPS)*, 2009.
- [24] H. Liu, M. Yang, Y. Gao, Y. Yin, Liang. Chen, Bilinear Discriminant Dictionary Learning for Face

- Recognition, *Pattern Recognition*, 47 (2014) 1835-1845.
- [25] Z. Feng, M. Yang, L. Zhang, Yan. Liu, D. Zhang, Joint discriminant dimensionality reduction and dictionary learning for face recognition, *Pattern Recognition*, 46 (2013) 2134-2143.
 - [26] C. Wei, Y. Chao, Y. Yeh, Y. Wang, Locality-sensitive dictionary learning for sparse representation based classification, *Pattern Recognition*, 46 (2013) 1277-1287.
 - [27] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition? in: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 471-478.
 - [28] L. Zhang, M. Yang, X. Feng, Y. Ma, D. Zhang, Collaborative Representation based Classification for Face Recognition, *arXiv: 1204.2358*, 2012.
 - [29] L. Zhang, W. Zhou, P. Chang, J. Liu, Z. Yan, T. Wang, F. Li, Kernel Sparse Representation-based Classifier, *IEEE Trans. Signal Process*, 60 (4) (2012) 1684-1695.
 - [30] S. Gao, I. W. Tsang, L. Chia, Sparse Representation with kernels, *IEEE Trans. Image Process*, 22 (2) (2013) 423-434.
 - [31] J. Yin, Z. Liu, Z. Jin, W. Yang, Kernel sparse representation based classification, *Neurocomputing*, 77 (2012) 120-129.
 - [32] S. Gao, I. W. Tsang, L.-T. Chia, Kernel sparse representation for image classification and face recognition, in *Proc. ECCV*, 2010, pp. 1-14.
 - [33] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, R. Chellappa, Design of non-linear kernel dictionaries for object recognition, *IEEE Trans. Image Process*, 22 (12) (2013) 5123-5135.
 - [34] L. Wei, F. Xu, J. Yin, Kernel locality-constrained collaborative representation based discriminant analysis, *Knowledge-Based Systems*, 70 (2014) 212-220.
 - [35] D. Donoho, For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_1 -norm Solution is also the Sparse Solution, *Commun. Pure Appl. Math.* 59 (6) (2006) 797-829.
 - [36] E. Candès, J. Romberg, T. Tao, Stable Signal Recovery from Incomplete and Inaccurate Measurements, *Commun. Pure Appl. Math.* 59 (8) (2006) 1207-1223.
 - [37] R. Tibshirani, Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58 (1996) 267-288.
 - [38] J. Yang, K. Yu, T. Huang, Supervised translation-invariant sparse coding, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. pp. 3517-3524.
 - [39] J. Yang, Z. Wang, Z. Lin, S. Cohen, T. Huang, Couple dictionary training for image super-resolution, *IEEE Trans. Image Process*, 21 (8) (2012) 3467-3478.
 - [40] J. Yang, Z. Wang, Z. Lin, X. Shu, T. Huang, Bilevel sparse coding for couple feature spaces, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. pp. 2360-2367.
 - [41] A. Martinez, R. Benavente, *The AR Face Database*, 1998.
 - [42] S.A. Nene, S.K. Nayar and H. Murase, Columbia object image library (COIL-20), Technical Report CUCS-005-96, Columbia University, 1996.
 - [43] J. Hull, A database for handwritten text recognition research, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16 (5) (1994) 550-554.
 - [44] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, expression database, *IEEE Trans. Pattern Anal. Mach. Intell.* 25(12) (2005) 1615-1618.
 - [45] G. Zhang, H. Sun, Z. Ji, Q. Sun, Label propagation based on collaborative representation for face recognition, *Neurocomputing*, 171 (2015) 1193-1204.