Joint Discriminative Dimensionality Reduction and Dictionary Learning for Face Recognition

Zhizhao Feng*, Meng Yang*, Lei Zhang[1], Yan Liu and David Zhang

Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China

Abstract: *In linear representation based face recognition (FR), it is expected that a discriminative dictionary can be learned from the training samples so that the query sample can be better represented for classification. On the other hand, dimensionality reduction is also an important issue for FR. It can not only reduce significantly the storage space of face images, but also enhance the discrimination of face feature. Existing methods mostly perform dimensionality reduction and dictionary learning separately, which may not fully exploit the discriminative information in the training samples. In this paper, we propose to learn jointly the projection matrix for dimensionality reduction and the discriminative dictionary for face representation. The joint learning makes the learned projection and dictionary better fit with each other so that a more effective face classification can be obtained. The proposed algorithm is evaluated on benchmark face databases in comparison with existing linear representation based methods, and the results show that the joint learning improves the FR rate, particularly when the number of training samples per class is small.*

Keywords: *Dictionary learning, face recognition, dimensionality reduction, collaborative representation*

*The first two authors contribute equally to this work.

[1] Corresponding author. Email: cslzhang@comp.polyu.edu.hk.

1. Introduction

Face recognition (FR) methods have been studied for over 30 years, and various techniques have been developed [1-8, 13] to handle different problems in face recognition, such as illumination, pose, occlusion and small sample size, etc. Face images usually have a high dimensionality, which makes the storage space high and increases the computational cost. In addition, the high dimensionality also decreases the discrimination of face images. Therefore, many dimensionality reduction techniques [2, 9, 10-12, 14] have been developed to reduce the dimension of face images and enhance the discriminative features. The representative dimensionality reduction methods include Principal Component Analysis (PCA) [9], Linear Discriminate Analysis (LDA) [10], Locality Preserving Projection (LPP) [2], etc. These so-called subspace analysis based FR methods are simple to apply; however, they are less effective to handle the expression and illumination changes. When the training samples are insufficient, the subspace learned by these methods will be much biased.

In the subspace based FR methods, often the nearest neighbor (NN) classifier and SVM are used for the classification. Recently, a new face classification scheme, i.e., the sparse representation based classification (SRC) [6], was proposed. In SRC, a query face image is encoded over the original training set with sparsity constraint imposed on the encoding vector. The training set acts as a dictionary to represent the testing samples as a sparse linear combination of its atoms. The classification is then performed by checking which class leads to the smallest reconstruction residual of the query sample. The SRC classifier shows very competitive performance, but its performance will drop much when the training samples per class are insufficient. It is also claimed in [6] that dimensionality reduction is no longer critical in the SRC scheme and random projection can achieve similar results to PCA and LDA when the dimensionality is high enough. Nonetheless, if a lower dimensionality is required, PCA and LDA will have clear advantage over random projection. Some works [14, 15] has been done to investigate the dimensionality reduction for SRC. For example, Zhang *et al*. [14] proposed an unsupervised learning method for dimensionality reduction in SRC, and it leads to higher FR rates than PCA and random

projection. This validates that a well designed dimensionality reduction method can benefit the sparse classification scheme.

In the meantime, there has been an increasing interest in learning a dictionary to represent the query image instead of using the original training samples. In FR, the original face images may contain some redundant information, noise or other trivial information that will obstruct the correct recognition. In [16], Yang *et al.* proposed a metaface learning (MFL) algorithm to represent the training samples by a series of "metafaces" learnt from each class. Based on the classical KSVD algorithm [17], in [18] a DKSVD algorithm was developed to code the query image and use the coding coefficients for classification. In [19], a supervised algorithm was proposed to learn a dictionary as well as a classifier for image classification tasks (e.g., digit recognition, texture classification). In [20], a class-dependent supervised simultaneous orthogonal matching pursuit scheme was developed to solve the dictionary learning problem while increasing the inter-class discrimination. Very recently, a Fisher discrimination dictionary learning algorithm [3] was developed for sparse representation based pattern classification, and it shows very competitive performance with other dictionary learning based pattern classification schemes.

The dimensionality reduction (DR) and dictionary learning (DL) are mostly studied as two independent problems in FR. Usually, DR is performed first to the training samples and the dimensionality reduced data are used for DL. However, the pre-learned DR projection may not preserve the best features for DL. Intuitively, the DR and DL processes should be jointly conducted for a more effective FR. To this end, we propose a joint discriminative DR and DL (JDDRDL) scheme to exploit more effectively and robustly the discriminative information of training samples. The goal is that the face image features from different classes can be effectively separated by a dictionary in a subspace, which are to be determined. In the proposed JDDRDL, an energy functional is defined and an iterative optimization algorithm is given to alternatively optimize the dictionary and projection matrix. From some initialization, in each iteration, for a fixed projection $\boldsymbol{P}$, the desired dictionary $\boldsymbol{D}$ can be updated; then with the updated dictionary $\boldsymbol{D}$, the projection matrix $\boldsymbol{P}$ can be refined. After several iterations, the learned $\boldsymbol{P}$ and $\boldsymbol{D}$ together can lead to a more effective FR system.

One important advantage of the proposed JDDRDL scheme is that it is more robust to the small sample size problem than state-of-art linear representation based face classification methods [3, 6, 14, 16]. The discriminative DR methods such as LDA and the linear representation based methods such as SRC usually require that the number of training samples per class cannot be too small, and their performance can be much reduced if the training sample is insufficient. By exploiting more effectively the discriminative information of training sample via learning the projection and dictionary simultaneously, the proposed JDDRDL shows more robust FR capability when the training sample size per class is small, for example 2~5 samples per class.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 presents in details the JDDRDL algorithm. Section 4 presents the experimental results; and Section 5 concludes the paper.

2. Related Work

2.1. PCA and LDA

As the most representative unsupervised DR method, PCA extracts the eigenvector of the high dimension data, and projects the high dimension data into a linear subspace spanned by leading eigenvectors, seeking a subspace with the maximized variance. PCA is very simple and efficient in reducing the sensitivity to Gaussian noise and some trivial information; however, PCA aims to preserve the global energy of face images but not the discrimination of face images. In contrast, as the most representative supervised DR method, LDA seeks directions which are best for discrimination. LDA finds projections that can minimize the variation of samples in the same class while maximizing the variation between different classes. LDA is effective for classification; however, it is sensitive to the number of training samples per class. In addition, the reduced dimensionality cannot be greater than the number of classes, which limits LDA's applications in practice.

2.2. SRC [6] and CRC (collaborative representation classification [26])

The SRC scheme proposed by Wright *et al*. [6] uses sparse representation for FR. Suppose $\boldsymbol{A}_k=[\boldsymbol{s}_{k,1}, \boldsymbol{s}_{k,2}, \ldots, \boldsymbol{s}_{k,n}] \in \Re^{m\times n}$ is the training dataset of the k^{th} class, where $\boldsymbol{s}_{k,j}$, j=1,2,…,n, is an m-dimensional vector stretched by the j^{th} sample of class k. For a test sample $\boldsymbol{y} \in \Re^m$ from class k, generally it can be well approximated as the linear combination of the samples from $\boldsymbol{A}_i$, i.e., $\boldsymbol{y} \approx \sum_{j=1}^{n} \alpha_{k,j}\boldsymbol{s}_{k,j} = \boldsymbol{A}_k\boldsymbol{\alpha}_k$, where $\boldsymbol{\alpha}_k = [\alpha_{k,1}, \alpha_{k,2}, ..., \alpha_{k,n}]^T \in \Re^n$ is the coding vector. Suppose we have K classes, and let $\boldsymbol{A}$=[$\boldsymbol{A}_1$, $\boldsymbol{A}_2$,…, $\boldsymbol{A}_K$], then the linear representation of $\boldsymbol{y}$ can be written in terms of all training samples as $\boldsymbol{y} \approx \boldsymbol{A}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1;\ldots,\boldsymbol{\alpha}_k;\ldots;\ \boldsymbol{\alpha}_K]=[0,\ldots,0,\alpha_{i,1},\alpha_{i,2},\ldots,\alpha_{i,n},0,\ldots,0]^T$. Clearly, the non-zero element in the coefficient vector could well encode the identity of the test image $\boldsymbol{y}$. In SRC [6], the l_1-minimization is used to solve the coding vector: $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 \right\}$, where λ is a scalar constant. Then classification is made by $\text{identity}(\boldsymbol{y}) = \arg\min_k \{e_k\}$, where $e_k = \left\| \boldsymbol{y} - \boldsymbol{A}_k\hat{\boldsymbol{\alpha}}_k \right\|_2$.

The SRC achieves interesting FR results; however, the use of l_1-minimization makes it computationally expensive. SRC and its many variants [5, 14, 16] emphasize the role of l_1-norm sparsity in the success of SRC. Very recently, Zhang *et al*. [26] pointed out that the success of SRC mainly comes from the collaborative representation of the query image by using all the training samples, but not the l_1-norm sparsity imposed on the coding vector. Based on this finding, Zhang *et al.* proposed the collaborative representation based classification (CRC), where the l_2-norm is used to regularize the coding coefficients: $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_2^2 \right\}$. It is shown that when the facial feature dimension is not much less than the number of training samples, CRC could achieve similar FR rates to SRC but the time complexity is enormously reduced.

2.3. DR and DL under the SRC framework

It is claimed in [6] that SRC is insensitive to feature extraction when the dimensionality is high enough; however, a well learned DR matrix can lead to a more accurate and stable recognition result. In [14], an orthogonal DR matrix $\boldsymbol{P}$ was learnt under the framework of sparse representation, and it achieves better performance than Eigenfaces and Randomfaces in the SRC scheme. Specifically, the matrix $\boldsymbol{P}$ is learnt via the following objective function based on Leave-One-Out scheme:

$$J_{\boldsymbol{P},\{\boldsymbol{\beta}_i\}} = \arg\min \left\{ \sum_{i=1}^{N} \left(\left\| \boldsymbol{P}\boldsymbol{z}_i - \boldsymbol{P}\boldsymbol{A}_i\boldsymbol{\beta}_i \right\|_F^2 + \lambda_1 \left\| \boldsymbol{\beta}_i \right\|_1 \right) + \lambda_2 \left\| \boldsymbol{A} - \boldsymbol{P}^T\boldsymbol{P}\boldsymbol{A} \right\|_F^2 \right\} \quad s.t.\ \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}$$

where N is the number of training samples, $\boldsymbol{z}_i$ is the i^{th} sample of the training set $\boldsymbol{A}$ and $\boldsymbol{A}_i$ is the set of training samples in $\boldsymbol{A}$ excluding $\boldsymbol{z}_i$. As can be seen from the above objective function, the projection matrix $\boldsymbol{P}$ preserves the energy of training set $\boldsymbol{A}$ while keeping the coding vector of each sample $\boldsymbol{z}_i$ sparse.

In SRC, the original training samples are used as the dictionary to represent the query sample. Intuitively, a more accurate and discriminative representation can be obtained if we could optimize a dictionary from the original training samples. In [16], Yang *et al.* proposed a "metaface" learning method, where a dictionary $\boldsymbol{D}_k = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p]$ of metafaces is learned from each class of training samples $\boldsymbol{A}_k$ under the sparse representation model via optimizing $J_{\boldsymbol{D}_k,\boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{D}_k,\boldsymbol{\Lambda}} \left\| \boldsymbol{A}_k - \boldsymbol{D}_k\boldsymbol{\Lambda} \right\|_F^2 + \lambda \left\| \boldsymbol{\Lambda} \right\|_1$ s.t. $\boldsymbol{d}_j^T\boldsymbol{d}_j = 1, j = 1, \ldots, p$. The metaface dictionary $\boldsymbol{D}_k$ and the associated coefficient matrix $\boldsymbol{\Lambda}$ are optimized alternatively. The final metaface dictionary $\boldsymbol{D}$ is formed by concatenating all the K dictionaries $\boldsymbol{D}_k$.

Though the metaface learning method [16] improves the representation power of the dictionary, it does not truly aim to increase the discrimination power of $\boldsymbol{D}$ in the objective function. Yang *et al.* [3] recently proposed a DL method, namely the Fisher discrimination dictionary learning (FDDL), which embeds the Fisher criterion in the objective function design. The FDDL scheme has two remarkable features. First, the dictionary atoms are learnt to associate the class labels so that the reconstruction residual from each class can be used in classification; second, the Fisher criterion is also imposed on the coding coefficients so that they carry discriminative information for classification. Since both the

reconstruction residual and coding coefficients are discriminative, a new classification scheme is then proposed in FDDL to fuse the two types of information for a more robust pattern recognition task.

3. Joint Learning Model for Dimensionality Reduction and Dictionary Learning

3.1. The modeling

In the related works introduced in Section 2, the DR and DL processes are handled separately. Usually, the DR projection matrix can be learnt first to reduce the dimensionality of training samples, and then DL is performed to learn a dictionary from the dimensionality reduced dataset. To more effectively use the discrimination information in the training set $\boldsymbol{A}$, we propose to learn the DR matrix $\boldsymbol{P}$ and the dictionary $\boldsymbol{D}$ jointly so that a more accurate classification can be achieved.

For the projection matrix $\boldsymbol{P}$, we expect that it could preserve the energy of $\boldsymbol{A}$ while making the different classes $\boldsymbol{A}_i$ more separable in the subspace defined by $\boldsymbol{P}$. To this end, we propose to learn an orthogonal projection matrix, which could maximize the total scatter of $\boldsymbol{A}$ and the between-class scatter of $\boldsymbol{A}$ simultaneously. For the dictionary $\boldsymbol{D}$, we expect that it is able to faithfully represent the dimensionality reduced dataset $\boldsymbol{PA}$, while making the samples from the same class close to each other in the space spanned by $\boldsymbol{D}$. With the above considerations, in this paper we propose the following joint discriminative dimensionality reduction and dictionary learning (JDDRDL) model to optimize $\boldsymbol{P}$ and $\boldsymbol{D}$:

$$J_{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}} = \arg\min_{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}} \left\{ \begin{array}{l} \sum_{k=1}^{K} \left(\left\| \boldsymbol{PA}_k - \boldsymbol{D}_k\boldsymbol{\Lambda}_k \right\|_F^2 + \lambda_1 \left\| \boldsymbol{\Lambda}_k \right\|_F^2 + \lambda_2 \left\| \boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k \right\|_F^2 \right) \\ -\gamma_1 \left\| \boldsymbol{PA}_t \right\|_F^2 - \gamma_2 \left\| \boldsymbol{PA}_b \right\|_F^2 \end{array} \right\} \text{ s.t. } \boldsymbol{d}_{k,j}^T \boldsymbol{d}_{k,j} = 1,\ \forall k, j,\ \boldsymbol{PP}^T = \boldsymbol{I} \quad (1)$$

where $\boldsymbol{D}_k$ is the sub-dictionary for class k and $\boldsymbol{D}$=[$\boldsymbol{D}_1$, $\boldsymbol{D}_2$, …, $\boldsymbol{D}_K$] forms the whole dictionary; $\boldsymbol{\Lambda}_k$ represents the coding coefficient matrix of $\boldsymbol{PA}_k$ over $\boldsymbol{D}_k$; $\boldsymbol{A}_t$ is the centralized training set, i.e., $\boldsymbol{A}_t = \boldsymbol{A}–\boldsymbol{M}$ with each column of $\boldsymbol{M}$ being the mean vector $\boldsymbol{m}$ of all samples in $\boldsymbol{A}$; $\boldsymbol{A}_b$ is the class-specific centralized dataset of $\boldsymbol{A}$, i.e., $\boldsymbol{A}_b$ = [$\boldsymbol{M}_1–\boldsymbol{M}$, …, $\boldsymbol{M}_K–\boldsymbol{M}$] with each column of $\boldsymbol{M}_k$ being the mean vector $\boldsymbol{m}_k$ of samples

in $\boldsymbol{A}_k$; $\boldsymbol{\Gamma}_k$ is a matrix with each column of it being the mean of the columns in $\boldsymbol{\Lambda}_k$; λ_1, λ_2, γ_1, and γ_2 are positive scalars. We require that each atom $\boldsymbol{d}_{k,j}$ in dictionary $\boldsymbol{D}_k$ has unit norm.

Let's make a more detailed look of the JDDRDL model in Eq. (1). By requiring that $\boldsymbol{P}$ is orthogonal, minimizing the term $-\|\boldsymbol{P}\boldsymbol{A}_t\|_F^2$ (i.e., maximizing $\|\boldsymbol{P}\boldsymbol{A}_t\|_F^2$) guarantees that the energy of $\boldsymbol{A}_t$ can be well preserved because we can reconstruct $\boldsymbol{A}_t$ by $\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{A}_t$. On the other hand, minimizing the term $-\|\boldsymbol{P}\boldsymbol{A}_b\|_F^2$ will enhance the discrimination between different classes after projection because it aims to maximize the distance between the class centers. Minimizing $-\|\boldsymbol{P}\boldsymbol{A}_t\|_F^2$ and $-\|\boldsymbol{P}\boldsymbol{A}_b\|_F^2$ simultaneously will also make the within class scatter of dataset $\boldsymbol{A}$ small.

By coding $\boldsymbol{P}\boldsymbol{A}_k$ over $\boldsymbol{D}_k$, we minimize the coding residual $\|\boldsymbol{P}\boldsymbol{A}_k - \boldsymbol{D}_k\boldsymbol{\Lambda}_k\|_F^2$ to ensure the representation power of dictionary $\boldsymbol{D}_k$. Note that we use the Frobenius norm, instead of the sparse l_1-norm, to regularize the coding coefficients by $\|\boldsymbol{\Lambda}_k\|_F^2$. This is based on the recent findings [26] that the l_1-norm sparsity does not play the key role in sparse representation based FR. However, using the Frobenius norm to regularize $\boldsymbol{\Lambda}_k$ significantly reduces the time complexity for optimization without sacrificing the performance. Finally, the minimization of $\|\boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k\|_F^2$ enforces the coding coefficients of the samples in class k to be close to their mean, reducing the variations of the coding vectors of each class. This term minimizes the within class scatter in the domain spanned by the dictionary $\boldsymbol{D}_k$.

Overall, in the JDDRDL model in Eq. (1), the targeted projection $\boldsymbol{P}$ and dictionary $\boldsymbol{D}$ will make the training samples have larger between class distances and smaller within class variations. Ideally, if $\boldsymbol{P}$ and $\boldsymbol{D}$ could be well optimized, more accurately classification of the query sample $\boldsymbol{y}$ can be obtained. Next, let's discuss how to do the minimization of Eq. (1).

3.2. The optimization

The JDDRDL objective function in Eq. (1) is non-convex. Like other authors have done when trying to solve similar optimization problems, here we use a two-stage alternative direction approach to solving it.

We partition the whole optimization into two sub-problems: fix the projection matrix $\boldsymbol{P}$ and solve for the dictionary $\boldsymbol{D}$ and the coefficient $\boldsymbol{\Lambda}$; and fix $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$ to update $\boldsymbol{P}$. These two sub-problems are solved alternatively and iteratively, and we stop at a good point to get the locally optimal solutions of $\boldsymbol{P}$ and $\boldsymbol{D}$. Because the algorithm could only get a local optimal solution, different initializations of $\boldsymbol{P}$ and $\boldsymbol{D}$ may result in different final solutions of $\boldsymbol{P}$ and $\boldsymbol{D}$. In our algorithm, we use PCA to initialize $\boldsymbol{P}$, and use the original training samples to initialize $\boldsymbol{D}$. Similar classification rates could be achieved if we initialize $\boldsymbol{D}$ randomly, though the resolved $\boldsymbol{D}$ will be different for different initializations. The whole optimization algorithm is presented in detail as follows.

Step 1) Initializing $\boldsymbol{P}$. We use PCA to initialize $\boldsymbol{P}$. That is, the initial $\boldsymbol{P}$ is the PCA transformation matrix of the training data $\boldsymbol{A}$.

Step 2) Fix $\boldsymbol{P}$, and solve $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$. In this case, the objective function in Eq. (1) reduces to

$$J_{\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}} = \arg\min_{\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}} \sum_{k=1}^{K} \left(\left\| \boldsymbol{P}\boldsymbol{A}_k - \boldsymbol{D}_k\boldsymbol{\Lambda}_k \right\|_F^2 + \lambda_1 \left\| \boldsymbol{\Lambda}_k \right\|_F^2 + \lambda_2 \left\| \boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k \right\|_F^2 \right) \text{ s.t. } \boldsymbol{d}_{k,j}^T \boldsymbol{d}_{k,j} = 1,\ \forall k, j \qquad (2)$$

Obviously, the above objective function can be partitioned as K individual problems, and we can optimize each pair $\{\boldsymbol{D}_k, \boldsymbol{\Lambda}_k\}$ separately as

$$J_{(\boldsymbol{D}_k,\boldsymbol{\Lambda}_k)}^{(k)} = \arg\min_{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k} \left(\left\| \boldsymbol{P}\boldsymbol{A}_k - \boldsymbol{D}_k\boldsymbol{\Lambda}_k \right\|_F^2 + \lambda_1 \left\| \boldsymbol{\Lambda}_k \right\|_F^2 + \lambda_2 \left\| \boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k \right\|_F^2 \right) \text{ s.t. } \boldsymbol{d}_{k,j}^T \boldsymbol{d}_{k,j} = 1,\ \forall j$$

$\boldsymbol{D}_k$ and $\boldsymbol{\Lambda}_k$ are also solved alternatively and iteratively. To make the optimization easier, we initialize $\boldsymbol{\Gamma}_k$ as zero, and in the following iterations $\boldsymbol{\Gamma}_k$ can be calculated as the column mean matrix of the updated coefficient matrix $\boldsymbol{\Lambda}_k$. Therefore, $\boldsymbol{\Gamma}_k$ can be viewed as a known constant matrix in optimizing $\boldsymbol{D}_k$ and $\boldsymbol{\Lambda}_k$ in each iteration.

From some initialization of $\boldsymbol{D}_k$ (for example, random initialization), the coding coefficients $\boldsymbol{\Lambda}_k$ can be computed. In each iteration, once $\boldsymbol{D}_k$ is given, we can readily have an analytical solution of $\boldsymbol{\Lambda}_k$ as follows:

$$\boldsymbol{\Lambda}_k = \left(\boldsymbol{D}_k^T \boldsymbol{D}_k + (\lambda_1 + \lambda_2)\boldsymbol{I} \right)^{-1} \left(\boldsymbol{D}_k^T \boldsymbol{P}\boldsymbol{A}_k + \lambda_2 \boldsymbol{\Gamma}_k \right) \qquad (3)$$

When $\boldsymbol{\Lambda}_k$ is obtained, the dictionary $\boldsymbol{D}_k$ can then be updated. The procedures of updating $\boldsymbol{D}_k$ are the same as those in [16].

After several iterations, all the $\boldsymbol{D}_k$ and $\boldsymbol{\Lambda}_k$ can be obtained, and we can consequently obtain the whole dictionary $\boldsymbol{D}$ and the associated coefficient matrix $\boldsymbol{\Lambda}$.

Step 3) Fix $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$, update $\boldsymbol{P}$. Let $\boldsymbol{X}=\boldsymbol{D}\boldsymbol{\Lambda}$, the objective function in Eq. (1) is reduced to:

$$J_{\boldsymbol{P}} = \arg\min_{\boldsymbol{P}} \left\{ \left\| \boldsymbol{P}\boldsymbol{A} - \boldsymbol{X} \right\|_F^2 - \gamma_1 \left\| \boldsymbol{P}\boldsymbol{A}_t \right\|_F^2 - \gamma_2 \left\| \boldsymbol{P}\boldsymbol{A}_b \right\|_F^2 \right\} \text{ s.t. } \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I} \tag{4}$$

The above sub-objective function $J_{\boldsymbol{P}}$ is itself non-convex, and we can have a local minimum of it as follows. First, since $\boldsymbol{P}\boldsymbol{P}^T=\boldsymbol{I}$, we have

$$\left\| \boldsymbol{P}\boldsymbol{A} - \boldsymbol{X} \right\|_F^2 = tr(\boldsymbol{P}\varphi(\boldsymbol{P})\boldsymbol{P}^T) \tag{5}$$

where $\phi(\boldsymbol{P}) = (\boldsymbol{A} - \boldsymbol{P}^T\boldsymbol{X})(\boldsymbol{A} - \boldsymbol{P}^T\boldsymbol{X})^T$. Let $\boldsymbol{S}_t=\boldsymbol{A}_t\boldsymbol{A}_t^T$ and $\boldsymbol{S}_b=\boldsymbol{A}_b\boldsymbol{A}_b^T$, we have $\left\| \boldsymbol{P}\boldsymbol{A}_t \right\|_F^2 = tr(\boldsymbol{P}\boldsymbol{S}_t\boldsymbol{P}^T)$ and $\left\| \boldsymbol{P}\boldsymbol{A}_b \right\|_F^2 = tr(\boldsymbol{P}\boldsymbol{S}_b\boldsymbol{P}^T)$. $J_{\boldsymbol{P}}$ can then be rewritten as

$$\begin{aligned} J_{\boldsymbol{P}} &= \arg\min_{\boldsymbol{P}} \left\{ tr\left(\boldsymbol{P}\varphi(\boldsymbol{P})\boldsymbol{P}^T\right) - \gamma_1 tr\left(\boldsymbol{P}\boldsymbol{S}_t\boldsymbol{P}^T\right) - \gamma_2 tr\left(\boldsymbol{P}\boldsymbol{S}_b\boldsymbol{P}^T\right) \right\} \quad \text{s.t. } \boldsymbol{P}\boldsymbol{P}^T=\boldsymbol{I} \\ &= \arg\min_{\boldsymbol{P}} tr\left(\boldsymbol{P}\left(\varphi(\boldsymbol{P}) - \gamma_1\boldsymbol{S}_t - \gamma_2\boldsymbol{S}_b\right)\boldsymbol{P}^T\right) \end{aligned} \tag{6}$$

To solve the above minimization in the current iteration h, we use $\varphi(\boldsymbol{P}_{(h\text{-}1)})$ to approximate the $\varphi(\boldsymbol{P})$ in Eq. (6), where $\boldsymbol{P}_{(h\text{-}1)}$ is the projection matrix obtained in iteration h-1. By using the Eigen Value Decomposition (EVD) technique, we have

$$\left[\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{U}\right] = EVD\left(\varphi(\boldsymbol{P}_{(h-1)}) - \gamma_1\boldsymbol{S}_t - \gamma_2\boldsymbol{S}_b\right) \tag{7}$$

where $\boldsymbol{\Sigma}$ is diagonal matrix formed by the eigenvalues of $\left(\varphi(\boldsymbol{P}_{(h-1)}) - \gamma_1\boldsymbol{S}_t - \gamma_2\boldsymbol{S}_b\right)$. Then we can take the updated $\boldsymbol{P}$ as the first l most important eigenvectors in $\boldsymbol{U}$, i.e., let $\boldsymbol{P}_{(h)}=\boldsymbol{U}(1{:}l, :)$. However, in this way the update of $\boldsymbol{P}$ may be too big, and make the optimization of the whole system in Eq. (1) unstable. Therefore, we choose to update $\boldsymbol{P}$ gradually in each iteration and let

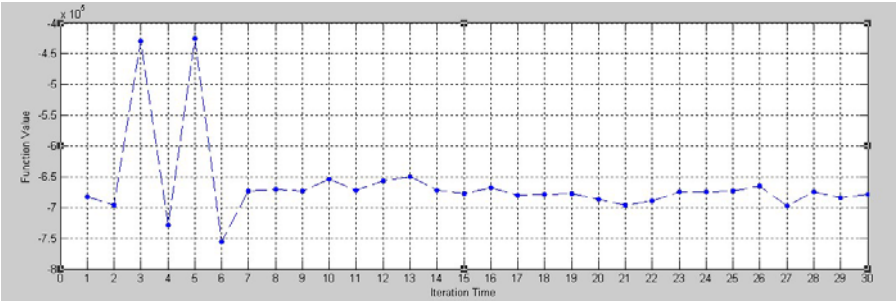
$$\boldsymbol{P}_{(h)}=\boldsymbol{P}_{(h\text{-}1)}+c(\boldsymbol{U}(1{:}l, :)-\boldsymbol{P}_{(h\text{-}1)}) \tag{8}$$

where c is a small positive constant to control the change of $\boldsymbol{P}$ in iterations.

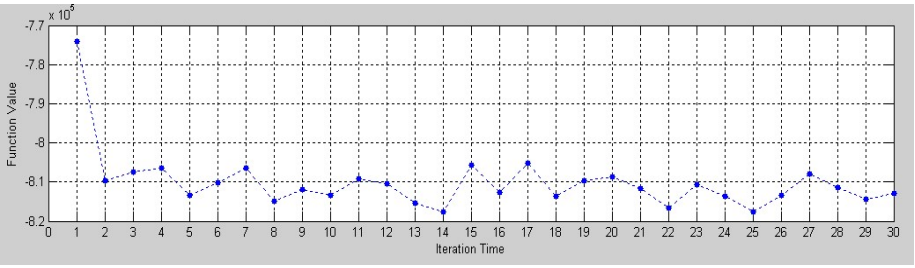
Step 4) Stopping criterion. If the maximum iteration number is reached, or the difference between the objective function $J_{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}}$ in adjacent iterations is smaller a preset value ε, then stop and output $\boldsymbol{P}$ and $\boldsymbol{D}$. Otherwise go back to Step 2.

3.3. Convergence of the JDDRDL model

The proposed JDDRDL model in Eq. (1) is jointly non-convex to the unknown variables, and thus the proposed optimization algorithm in Section 3.2 can at most reach a local minimum of it. In Step 2, the sub-problem is convex to each of $\{\boldsymbol{D}_k, \boldsymbol{\Lambda}_k\}$ when the other one is fixed, and our algorithm will lead to a local minimum of this sub-problem. However, in Step 3, Eq. (6) is an approximate formulation to the original sub-problem in Eq. (4), and thus the obtained solution is only an approximation to the local minimum of the sub-problem. Overall, the convergence of our algorithm cannot be guaranteed but by experience we can have a stable solution.



(a) AR database



(b) MPIE database

Figure 1: The convergence curves of JDDLDR algorithm on the (a) AR and (b) MPIE databases. The parameter values are $\lambda_1 = \lambda_2 = 0.005$, $\gamma_1 = 10$ and $\gamma_2 = 1$.

Let's use the AR database [27] and MPIE database [28] as examples to illustrate the optimization process of JDDRDL. The dimensionality of the face images is reduced to 300. The curves of the objective function $J_{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}}$ vs. the iteration number are plotted in Fig. 1(a) and Fig. 1(b), respectively, for the two databases. We can see that after several iterations (e.g., 6 iterations), the value of the objective function becomes stable, and it varies only in a small range. Usually, the iteration will stop within 15 times. Our experimental results also show that stopping the minimization with more or less iterations, the resulted projection $\boldsymbol{P}$ and dictionary $\boldsymbol{D}$ will lead to almost the same FR rates. This indicates that although the proposed JDDRDL algorithm cannot lead to an ideal convergence, it is not sensitive to the iteration number. In our experiments, we set the maximal iteration number as 15 and it works well.

3.4. The classification scheme

After we obtain the projection matrix $\boldsymbol{P}$, the query sample $\boldsymbol{y}$ can be projected into the lower dimensional space by $\boldsymbol{Py}$, and then the lower dimensional feature $\boldsymbol{Py}$ can be coded over the dictionary $\boldsymbol{D}$. Here we adapted the collaborative representation model with l_2-norm regularization [26] for coding because of its effectiveness and efficiency:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{Py} - \boldsymbol{D\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_2^2 \right\} \tag{9}$$

where λ is a positive scalar. Obviously, we have $\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{D}^T\boldsymbol{Py}_0$. The resulted coding vector can be written as $\hat{\boldsymbol{\alpha}}$ =[$\hat{\boldsymbol{\alpha}}_1$; …; $\hat{\boldsymbol{\alpha}}_k$; …; $\hat{\boldsymbol{\alpha}}_K$], where $\hat{\boldsymbol{\alpha}}_k$ is the sub-coding-vector associated with each sub-dictionary $\boldsymbol{D}_i$.

Once the coding vector $\hat{\boldsymbol{\alpha}}$ is computed, the classification can be conducted based on the reconstruction residual of each class, as that in SRC [6] or CRC [26]. However, in the proposed JDDRDL algorithm, the mean of the coding vectors $\boldsymbol{\Lambda}_k$ of each class, denoted by $\boldsymbol{u}_k$, is also learned, and the distance between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{u}_k$ is also useful for classification, as shown in [3]. Therefore, we adopted the classifier in [3] for the final classification. Let

$$e_k = \left\| \boldsymbol{P}\boldsymbol{y} - \boldsymbol{D}_k \hat{\boldsymbol{\alpha}}_k \right\|_2^2 + \omega \left\| \hat{\boldsymbol{\alpha}}_k - \boldsymbol{\mu}_k \right\|_2^2 \tag{10}$$

where ω is the constant to balance the contribution of the two terms. The final classification is performed by $\text{identity}(\boldsymbol{y}) = \arg\min_k \{e_k\}$.

4. Experimental results

In this section, we use several benchmark face recognition databases to verify the performance of our proposed JDDRDL scheme. The representative algorithms that employs dictionary learning and/or dimensionality reduction under the SRC framework, including SRC [6] with PCA and LDA, CRC [26] with PCA and LDA, metaface learning for SRC (MFL-SRC) [16], dimension reduction for SRC (DR-SRC) [14] and the recently proposed FDDL [3], are used for comparison. The l_1_*ls* [21] toolbox, which is a stable l_1-minimization solver, is used to solve the l_1-minimization problem (for other l_1-minimization solvers, please see [22-25]) in the SRC related algorithms. On each database, we first test the robustness of these competing methods to the number of training samples, and then show their results with different dimensionalities of the features.

4.1 Parameter selection

There are four parameters, λ_1, λ_2, γ_1 and γ_2, in our JDDLDR model in Eq. (1). The four parameters have very clear physical meaning, which could guide the setting of these parameters. (λ_1, λ_2) are to update the dictionary $\boldsymbol{D}_k$ and coding coefficient $\boldsymbol{\Lambda}_k$, while (γ_1, γ_2) are to update the projection matrix $\boldsymbol{P}$ for dimension reduction. Therefore in parameter selection, we could determine (λ_1, λ_2), and then determine (γ_1, γ_2). From Eq. (3), we can see that the setting of $\lambda_1$and λ_2 could simultaneously regularize the coding coefficient $\boldsymbol{\Lambda}_k$ and introduce discrimination via minimizing the within-class scatter of $\boldsymbol{\Lambda}_k$. Since each atom (i.e., the column vector) of $\boldsymbol{D}_k$ has a unit l_2-norm, we set $\lambda_1 = \lambda_2 = 0.005$ based on our experimental experience.

Parameters γ_1 and γ_2 are related to the learning of dimensionality reduction projection matrix $\boldsymbol{P}$. They should be set bigger compared with $\lambda_1$and λ_2, since trivial solutions (e.g., $\boldsymbol{P} \approx \text{Null}(\boldsymbol{A})$, i.e., $\boldsymbol{PA}\approx\boldsymbol{0}$) would be got if only the first three terms in Eq. (1) work. By experimental experience, we set γ_1=10 and γ_2=1 to mainly maximize the total scatter of the training samples, while introducing some discrimination between classes. In the testing stage, the scalar λ (refer to Eq. (9)) is set to 0.001 and ω (refer to Eq. (10)) is set to 0.01 in all experiments by experience.

4.2. FR results

a) AR database: The AR database [27] consists of over 4,000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separate sessions. In our experiments, a subset that contains 50 males and 50 females with 6 illumination and 8 expression variations in two sessions is used (please refer to Fig. 2 for some examples).

We randomly chose 2~7 samples per subject for training, while the other samples were used as query samples, all the samples were projected into a 550 dimensional subspace (Samples in LDA+SRC and LDA+CRC schemes were projected into a 99 dimensional subspace). The experiments were repeated 50 times to calculate the average recognition rate and the corresponding standard deviation. The FR rates by competing methods are listed in Table 1. It can be seen that when the number of training samples per class is not very small, e.g., 7 samples per class, the recognition rates by all competing methods are quite similar and satisfying. With the decrease of the number of training samples, the recognition rates of all methods drop, especially for LDA+SRC and LDA+CRC. This is mainly because LDA is sensitive to the number of training samples. The proposed JDDRDL achieves the highest FR rates among all the competing methods. Particularly, it is less sensitive to the small sample size problem. When the number of training samples per class is relatively high such as 6 or 7 samples per class, JDDRDL has very close recognition rates to FDDL. However, when the number of training samples are relatively low such as 2~5

samples per class, the difference between the recognition rates of JDDLDR and other methods is getting higher. Overall, JDDRDL's performance is very stable.

Figure 2: Some samples from the AR database.

Table 1: Recognition rates on the AR database with different number of training samples.

No. of training Samples	2	3	4	5	6	7
JDDRDL	0.734±0.037	0.759±0.026	0.818±0.020	0.897±0.017	0.929±0.020	0.941±0.022
DR-SRC	0.711±0.034	0.740±0.028	0.798±0.022	0.871±0.020	0.908±0.021	0.930±0.025
MFL-SRC	0.714±0.031	0.736±0.023	0.790±0.018	0.872±0.024	0.909±0.027	0.932±0.019
PCA+SRC	0.705±0.029	0.731±0.024	0.794±0.014	0.872±0.018	0.910±0.020	0.932±0.018
LDA+SRC	0.494±0.044	0.534±0.033	0.718±0.020	0.859±0.014	0.892±0.027	0.914±0.024
PCA+CRC	0.708±0.030	0.737±0.028	0.788±0.019	0.874±0.021	0.910±0.018	0.930±0.020
LDA+CRC	0.491±0.029	0.534±0.031	0.714±0.028	0.859±0.019	0.890±0.022	0.912±0.015
FDDL	0.690±0.032	0.702±0.029	0.796±0.015	0.888±0.020	0.924±0.022	0.933±0.028

Table 2: Recognition rates on the AR database under different feature dimensions.

Dimension	99	350	400	450	500	550
JDDLDR	---	0.805±0.018	0.813±0.023	0.823±0.021	0.822±0.027	0.818±0.020
DR-SRC	---	0.787±0.022	0.791±0.020	0.801±0.024	0.804±0.031	0.798±0.022
MFL-SRC	---	0.788±0.020	0.789±0.014	0.809±0.018	0.798±0.021	0.790±0.018
PCA+SRC	---	0.782±0.027	0.783±0.014	0.804±0.017	0.800±0.025	0.794±0.014
LDA+SRC	0.718±0.020	---	---	---	---	---
PCA+CRC	---	0.784±0.027	0.787±0.020	0.800±0.020	0.793±0.036	0.788±0.019
LDA+CRC	0.714±0.028	---	---	---	---	---
FDDL	---	0.782±0.023	0.794±0.019	0.802±0.024	0.801±0.034	0.796±0.015

We then evaluate the performance of JDDRDL on different dimensionalities. Four samples of each subject are randomly chosen for training, and all the remaining images are used as query images. The recognition rates with different feature dimensions by the competing methods are shown in Table 2. JDDLDR surpasses other competing schemes on average. It can be seen that when the dimensionality is

relatively low, e.g., 350, all the methods (except for LDA+SRC and LDA+CRC) have similar results. With the increase of feature dimension, e.g., above 450, the proposed JDDLDR shows visible improvement over the other methods.

b) Multi PIE database: The CMU Multi-PIE database [28] contains image of 337 subjects captured in four sessions with simultaneous variations in pose, expression, and illumination. Among these 337 subjects, all the 249 subjects in Session 1 are used (see Fig. 3 for example samples). We randomly selected 2 to 7 samples per subject as our training set while the other images were used as query set, and projected into a subspace of 550 dimensions (Samples in LDA+SRC and LDA+CRC schemes are projected into a subspace of 248 dimensions). Also, all experiments were repeated for 50 times to calculate the mean and standard deviation of the FR rates. Table 3 shows the results by different methods. We can draw similar conclusions to those on the AR database, i.e., the proposed JDDRDL achieves the best FR rates and its advantage over the other methods is more remarkable when the number of training samples is less sufficient.

Table 4 lists the recognition rates of the competing methods with different dimensions of features. Four images were randomly chosen from each subject for training set, and the remaining samples were used as for testing, and such experiments were repeated 50 times as well. Similar to what we observed on the AR database, JDDRDL achieves more remarkable improvement over the other methods with the increase of dimensionality. It is also noticed that LDA+SRC and LDA+CRC have good performance on MPIE since MPIE is a large scale dataset with 249 classes, which allows LDA to use enough number of projections to classify the query samples.

Figure 3: Some samples from the MPIE database.

Table 3: Recognition rates on the MPIE database with different number of training samples.

No. of training samples	2	3	4	5	6	7
JDDRDL	0.756±0.044	0.837±0.029	0.900±0.018	0.906±0.020	0.910±0.011	0.912±0.008
DR-SRC	0.744±0.047	0.824±0.033	0.876±0.024	0.888±0.022	0.902±0.025	0.904±0.010
MFL-SRC	0.741±0.034	0.826±0.020	0.871±0.021	0.881±0.013	0.889±0.012	0.907±0.014
PCA+SRC	0.743±0.039	0.822±0.040	0.880±0.029	0.891±0.028	0.894±0.009	0.905±0.016
LDA+SRC	0.421±0.040	0.795±0.026	0.874±0.020	0.884±0.016	0.895±0.014	0.910±0.009
PCA+CRC	0.745±0.037	0.820±0.033	0.875±0.015	0.893±0.030	0.898±0.013	0.907±0.013
LDA+CRC	0.414±0.042	0.801±0.028	0.877±0.026	0.880±0.019	0.900±0.020	0.908±0.019
FDDL	0.659±0.035	0.810±0.041	0.888±0.017	0.904±0.026	0.908±0.016	0.910±0.015

Table 4: Recognition rates on the MPIE database under different feature dimensions.

Dimension	248	350	400	450	500	550
JDDLDR	---	0.866±0.016	0.872±0.011	0.878±0.014	0.886±0.016	0.900±0.018
DR-SRC	---	0.858±0.015	0.867±0.017	0.864±0.010	0.875±0.020	0.876±0.024
MFL-SRC	---	0.853±0.011	0.859±0.010	0.865±0.016	0.871±0.017	0.871±0.021
PCA+SRC	---	0.870±0.014	0.874±0.021	0.867±0.021	0.878±0.018	0.880±0.029
LDA+SRC	0.874±0.020	---	---	---	---	---
PCA+CRC	---	0.873±0.013	0.874±0.014	0.870±0.019	0.877±0.019	0.875±0.015
LDA+CRC	0.877±0.026	---	---	---	---	---
FDDL	---	0.866±0.011	0.871±0.012	0.872±0.014	0.881±0.016	0.888±0.017

c) Extended Yale B Database: The extended Yale B [29] database contains about 2,414 frontal face images of 38 individuals taken under varying illumination conditions. We randomly chose 2 to 7 images from each person as training set, and used the rest images as testing set. Similarly, all the samples were projected into a subspace of 550 dimensions (Samples in LDA+SRC and LDA+CRC schemes are projected into a subspace of 37 dimensions) and the experiments were repeated 50 times. The FR results are shown in Table 5.

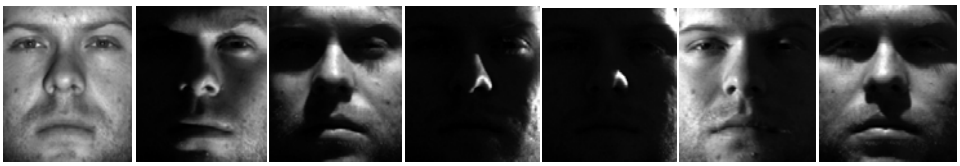


Figure 4: Some samples from the Extended Yale B database.

Table 5. Recognition rates on the Yale B database with different number of training samples.

No. of training samples	2	3	4	5	6	7
JDDRDL	0.549±0.034	0.653±0.036	0.674±0.025	0.682±0.022	0.696±0.030	0.705±0.024
DR-SRC	0.530±0.038	0.636±0.031	0.656±0.030	0.671±0.025	0.689±0.023	0.698±0.021
MFL-SRC	0.534±0.029	0.631±0.025	0.657±0.026	0.668±0.023	0.690±0.023	0.692±0.018
PCA+SRC	0.535±0.031	0.641±0.034	0.652±0.024	0.670±0.029	0.687±0.024	0.690±0.031
LDA+SRC	0.462±0.032	0.532±0.031	0.603±0.028	0.665±0.030	0.681±0.019	0.681±0.022
PCA+CRC	0.532±0.028	0.644±0.024	0.650±0.022	0.671±0.025	0.685±0.024	0.692±0.025
LDA+CRC	0.460±0.039	0.535±0.033	0.609±0.031	0.662±0.028	0.679±0.020	0.682±0.014
FDDL	0.441±0.042	0.538±0.037	0.636±0.023	0.675±0.021	0.693±0.017	0.701±0.025

Table 6. Recognition rates on the Extended Yale B database under different feature dimensions.

Dimension	37	350	400	450	500	550
JDDLDR	---	0.658±0.017	0.660±0.015	0.665±0.021	0.666±0.031	0.674±0.025
DR-SRC	---	0.644±0.019	0.647±0.017	0.648±0.022	0.651±0.028	0.656±0.030
MFL-SRC	---	0.640±0.022	0.640±0.025	0.642±0.029	0.645±0.033	0.657±0.026
PCA+SRC	---	0.640±0.026	0.641±0.019	0.644±0.018	0.650±0.026	0.652±0.024
LDA+SRC	0.603±0.028	---	---	---	---	---
PCA+CRC	---	0.637±0.014	0.645±0.022	0.649±0.023	0.652±0.024	0.650±0.022
LDA+CRC	0.609±0.031	---	---	---	---	---
FDDL	---	0.614±0.019	0.616±0.024	0.618±0.025	0.624±0.028	0.636±0.023

We then randomly selected 4 images from each subject as the training set, and took the remaining samples as the testing set. The FR rates under different dimensions are shown in Table 6. Compared with the AR database, the extended Yale B database has less expression variations but larger illumination changes (please see Fig. 4 for examples). When the number of training samples is insufficient, the FR becomes very challenging due to the large variation in illumination. From Tables 5 and 6, one can see that the proposed JDDLDR method achieves the highest recognition rates among the competing schemes. When there are only 3 training samples per subject, JDDLDR achieves about 10% higher recognition rate than FDDL, which is a state-of-the-art discriminative dictionary learning method. This is because FDDL performs dimensionality separately from the discriminative dictionary learning process so that it requires enough training samples to stably compute the statistics. By coupling the dimensionality reduction and dictionary learning processes, the proposed JDDLDR can

increase much the robustness to the number of training samples while yielding a discriminative dictionary.

d). FERET database: A pose subset of the FERET database [30] is used here, which includes the frontal face images marked with "*ba*", "*bj*", and "*bk*". Since there are only three samples for each subject, in each experiment we use two samples for training and the other one for testing. In the first experiment, the image marked with "*ba*" and "*bj*" were used as training samples, and totally 200 classes and 400 samples were used in the training set. The testing set includes the images marked with "*bk*" for each subject (please refer to Fig. 5 for examples). The FR result is shown in Table 7(a). In experiment 2, the images marked with "*ba*" and "*bk*" were used for training and "*bj*" was used for testing. The result is list in Table 7(b). Similarly, in the third experiment, images "*bj*" and "*bk*" were used for training and "*ba*" was used for testing. The result is list in the Table 7(c).

Figure 5: Some samples from the FERET database

Table 7(a). Recognition rates on the FERET database under different feature dimensions.

Dimension	199	350	400	450	500	550
JDDRDL	---	0.795	0.810	0.800	0.785	0.780
DR-SRC	---	0.790	0.790	0.785	0.785	0.775
MFL-SRC	---	0.795	0.795	0.785	0.770	0.770
PCA+SRC	---	0.785	0.785	0.790	0.785	0.775
LDA+SRC	0.715	---	---	---	---	---
PCA+CRC	---	0.790	0.795	0.790	0.785	0.775
LDA+CRC	0.715	---	---	---	---	---
FDDL	---	0.720	0.725	0.725	0.715	0.715

Table 7(b). Recognition rates on the FERET database under different feature dimensions.

Dimension	199	350	400	450	500	550
JDDRDL	---	0.895	0.900	0.900	0.895	0.895
DR-SRC	---	0.880	0.880	0.875	0.875	0.875
MFL-SRC	---	0.875	0.875	0.875	0.880	0880
PCA+SRC	---	0.890	0.890	0.890	0.885	0.880
LDA+SRC	0.730	---	---	---	---	---
PCA+CRC	---	0.890	0.895	0.890	0.890	0.885
LDA+CRC	0.735	---	---	---	---	---
FDDL	---	0.790	0.795	0.795	0.800	0.800

Table 7(c). Recognition rates on the FERET database under different feature dimensions.

Dimension	199	350	400	450	500	550
JDDRDL	---	0.915	0.930	0.940	0.920	0.920
DR-SRC	---	0.895	0.905	0.920	0.910	0.910
MFL-SRC	---	0.895	0.900	0.915	0.910	0.905
PCA+SRC	---	0.895	0.900	0.910	0.910	0.905
LDA+SRC	0.755	---	---	---	---	---
PCA+CRC	---	0.900	0.905	0.910	0.905	0.905
LDA+CRC	0.750	---	---	---	---	---
FDDL	---	0.790	0.800	0.815	0.810	0.810

Similar to the results in other databases, from Tables 7(a), 7(b) and 7(c) we see that the proposed JDDLDR achieves the highest recognition results in the three experiments, which demonstrates its capability to handle the small sample size problem. The LDA+SRC, LDA+CRC and FDDL methods do not work well on this dataset because their sensitivity to the number of training samples. Compared with DR-SRC, MFL-SRC, PCA+CRC and PCA+SRC, the JDDLDR can always achieve certain improvement in the three experiments.

e) FRGC database: FRGC version 2.0 [31] is a large-scale face database established under uncontrolled indoor and outdoor settings. Some examples are shown in Fig. 6. We use a subset (316 subjects having no less than 10 samples and 7318 images in total) of query face image database, which has large lighting, accessory (e.g., glasses), expression variations and image blur, etc. We randomly

chose 2 to 5 samples per subject as the training set, with the remaining as the testing set. The images were cropped to 32×42 and all the experiments were run 50 times to calculate the mean and standard deviation. Table 8 lists the face recognition results of all the competing methods with 400-diemsion projected feature (here the dimension of LDA feature is 315). It can be seen that the mean recognition rate of JDDRDL outperforms about 2% all the other methods in all cases. It clearly validates that both the learned projection matrix and the dictionary of JDDRDL could bring benefit to the final classification. Table 9 shows the recognition rates versus different feature dimensions when the number of training samples per subject is 3. We could draw similar conclusion, i.e., the proposed JDDRDL method performs the best in every dimension with almost 2% improvement over other methods.

Figure 6: Some samples from the FRGC 2.0 database

Table 8. Recognition rates on the FRGC 2.0 database with different number of training samples.

No. of training Samples	2	3	4	5
JDDRDL	0.7465±0.0095	0.8558±0.0065	0.9080±0.0044	0.9356±0.0030
DR-SRC	0.6979±0.0035	0.8005±0.0082	0.8622±0.0069	0.8910±0.0049
MFL-SRC	0.7147±0.0042	0.8294±0.0066	0.8863±0.0028	0.9133±0.0027
PCA+SRC	0.6968±0.0075	0.8003±0.0058	0.8572±0.0068	0.8893±0.0044
LDA+SRC	0.6319±0.0695	0.8211±0.0244	0.8738±0.0173	0.9126±0.0039
PCA+CRC	0.7322±0.0087	0.8346±0.0066	0.8879±0.0044	0.9162±0.0037
LDA+CRC	0.6630±0.0695	0.8362±0.0203	0.8814±0.0177	0.9160±0.0032
FDDL	0.732 ±0.0037	0.8317±0.0085	0.8842±0.0056	0.9124±0.0041

Table 9. Recognition rates on the FRGC 2.0 database under different feature dimensions.

Dimension	315	350	400	450	500
JDDLDR	---	0.8533±0.0071	0.8558±0.0065	0.8616±0.0068	0.8589±0.0052
DR-SRC	---	0.7964±0.0082	0.8005±0.0082	0.8018±0.0088	0.8028±0.0082
MFL-SRC	---	0.8297±0.0102	0.8294±0.0066	0.8300±0.0074	0.8311±0.0075
PCA+SRC	---	0.7979±0.006	0.8003±0.0058	0.8015±0.006	0.8024±0.0058
LDA+SRC	0.8211±0.0244	---	---	---	---
PCA+CRC	---	0.8309±0.0044	0.8346±0.0066	0.8386±0.0046	0.8407±0.0052
LDA+CRC	0.8362±0.0203	---	---	---	---
FDDL	---	0.8279±0.0084	0.8317±0.0085	0.8320±0.0076	0.8340±0.0074

4.3. Statistical significance test

In order to more convincingly show the effectiveness of the proposed method, we perform statistical significance test to verify whether the improvement of JDDLDR over other methods is significant. More specifically, we perform a *t-test* (please refer to page 155 of [32]) on the null hypothesis that the improvement of JDDLDR over some competing method X is insignificant (i.e., the difference of the recognition rates between JDDLDR and X come from distributions with mean less than zero) by using all the recognition rates in each experiment.

We focus on two outputs of the statistical significance test: H and P. H=0 indicates that the null hypothesis can not be rejected at some significance level, and P is the probability of observing the result of H=0 (small values of P cast doubt on the validity of the null hypothesis). When setting the significance level as 1%, we get that H=1 (i.e., the proposed JDDLDR is significantly better than all the other methods) holds well in almost all comparisons except for the cases of FDDL (with P=5.36%), LDA+CRC (with P=9.99%) and LDA+SRC (with P=8.75%) in MPIE. The reason is that the performance of FDDL approaches to JDDLDR when the number of training sample is large, and the dimensionality of JDDLDR is not large enough (e.g., 350, 400 and 450) when comparing with LDA+CRC/SRC. However, when the number of training samples per subject is small (e.g., 2 or 3), the

advantage of JDDLR over FDDL is significant (e.g., about 10% improvement when the number of training samples per class is 2), and when the dimensionality is large enough, JDDLDR could also outperform LDA+CRC/SRC. With setting the significance level as 10%, we get H=1 in all cases, validating that JDDLDR significantly outperforms all the other methods in all cases.

5. Conclusion

In this paper we proposed a joint discriminative dimensionality reduction and dictionary learning (JDDLDR) scheme for face recognition. Unlike many methods which focus on dictionary learning (DL) and use PCA or LDA for dimensionality reduction (DR), JDDLDR considers the interaction between DR and DL procedures by coupling them into a unified framework for energy minimization. The DR matrix projects the data into a lower dimensional subspace where the total scatter and between-class scatter of the training data are maximized, while the learned dictionary associated with the DR matrix is ensured to have a strong representative ability. In classification, both the representation residual and the distance between the coding vector and the mean vector of each class were considered. The experimental results on representative face databases demonstrated that the proposed JDDRDL method surpasses many state-of-the-arts face recognition methods.

6. References

[1] H.-C. Kim, D. Kim, and S. Y. Bang, Face recognition using LDA mixture model, *Pattern Recognition Letters* 24(15):2815-2821, 2003.

[2] X. He, P. Niyogi, Locality preserving projections, in: *Proceedings of the NIPS*, 2003.

[3] M. Yang, L. Zhang, X. Feng, D. Zhang, Fisher discrimination dictionary learning for sparse representation, in: *Proceedings of the ICCV*, 2011.

[4] I. Craw, N. Costen, T. Kato, S. Akamatsu, How should we represent faces for automatic recognition, *IEEE Transaction on Pattern Analysis and Machine Intelligence,* 21(8): 725-736, 1999.

[5] A.Wagner, J.Wright, A. Ganesh, Z. Zhou, Y. Ma, Towards a practical face recognition system: Robust registration and illumination by sparse representation, in: *Proceeding of the CVPR*, 2009

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(2): 210–227, 2009.

[7] S. Li, X. Hou, H. Zhang, Q. Cheng, Learning spatially localized, parts-based representation, in: *Proceeding of the CVPR*, 2001.

[8] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: *Proceeding of the NIPS*, 2001.

[9] M. Turk, A. Pentland, Face recognition using eigenfaces, in: *Proceeding of the CVPR*, 1991.

[10] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7):711-720, 1997.

[11] J. Ham, D. Lee, S. Mika, B. Scholkopf, A kernel view of the dimensionality reduction of manifolds, in: *Proceeding of the ICML*, 2004.

[12] W. Zhao, R. Chellappa, P.J. Phillips, Subspace linear discriminate analysis for face recognition, *Technical Report CAR-TR-914, Center for Automation Research, Univ. of Maryland*, 1999.

[13] W. Zhao, R. Chellappa, J. Phillips, A. Rosenfeld, Face recognition: A literature survey, *ACM Computing Survey,* 35(4): 399–458, 2003.

[14] L. Zhang, M. Yang, Z. Feng, D. Zhang, On the dimensionality reduction for sparse representation based face recognition, in: *Proceeding of the ICPR*, 2010

[15] Yen-Yu Lin, Tyng-Luh Liu, Chiou-Shann Fuh, Multiple kernel learning for dimensionality reduction, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 33(6):1147 – 1160, 2011.

[16] M. Yang, L. Zhang, J. Yang, D. Zhang, Metaface learning for sparse representation based face recognition, in: *Proceeding of the ICIP*, 2010.

[17] M. Aharon, M. Elad, A. Bruckstein, The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation, *IEEE Transaction on Signal Processing*, 51(11): 4311-4322, 2006.

[18] Q. Zhang, B. Li, Discriminative K-SVD for dictionary learning in face recognition, in: *Proceeding of the CVPR*, 2010.

[19] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Supervised dictionary learning, in: *Proceeding of the NIPS*, 2008.

[20] F. Rodriguez, G. Sapiro, Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries, *Technical report, University of Minnesota, IMA Preprint 2213*, 2007.

[21] S. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, A interior-point method for large-scale l_1-regularized least squares, *IEEE Journal on Selected Topics in Signal Processing*, 1 (4):606–617, 2007.

[22] D. Malioutove, M. Cetin, A. Willsky, Homotopy continuation for sparse signal representation, in: *Proceeding of the ICASS*, 2005.

[23] A. Yang, S. Sastry, A. Ganesh, Y. Ma, Fast l_1-minimization algorithms and application in robust face recognition: A review, in: *Proceeding of the ICIP*, 2010

[24] A. Beck, M. Teboulle, A fast iterative shrinkagethresholding algorithm for linear inverse problems, *SIAM Journal on Image Science*, 2(1):183–20, 2009.

[25] M. Figueiredo, R. Nowak, S. Wright, Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems, *IEEE Journal on Selected Topics in Signal Processing*, 1(4):586–597, 2007.

[26] L. Zhang, M. Yang, X. Feng, Sparse Representation or Collaborative Representation: Which Helps Face Recognition, in: *Proceeding of the ICCV*, 2011.

[27] A.Martinez, R. Benavente, The AR face database, *CVC Technical Report 24*, 1998.

[28] R. Gross, I. Matthews. J. Cohn, T. Kanade, S. Baker, Multi-PIE, *Image and Vision Computing*, 28:807–813, 2010.

[29] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and pose, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[30] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, The FERET evaluation methodology for face recognition algorithms, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(10):1090 – 1104, 2000.

[31] P. J. Phillips, P. J. Flynn, W. T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. J. Worek, Overiew of the face recognition grand challenge, in: *Proceeding of the CVPR*, 2005.

[32] R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing (Second Edition)*, ELSEVIER Academic Press, 2005.