Pattern Recognition 81 (2018) 341–356

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Robust, discriminative and comprehensive dictionary learning for face recognition

Guojun Lin[a,b], Meng Yang[a,e,*], Jian Yang[c], Linlin Shen[d], Weicheng Xie[d]

[a] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
[b] School of Automation and Electric Information, Sichuan University of Science and Engineer, Zigong, China
[c] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China
[d] School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[e] Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-Sen University, Ministry of Education, China

ARTICLE INFO

Article history:
Received 31 October 2016
Revised 7 February 2018
Accepted 23 March 2018
Available online 30 March 2018

Keywords:
Dictionary learning
Face recognition
Sparse representation

ABSTRACT

For sparse representation or sparse coding based image classification, the dictionary, which is required to faithfully and robustly represent query images, plays an important role on its success. Learning dictionaries from the training data for sparse coding has shown state-of-the-art results in image classification and face recognition. However, for face recognition, conventional dictionary learning methods cannot well learn a reliable and robust dictionary due to suffering from the small-sample-size problem. The other significant issue is that current dictionary learning do not completely cover the important components of signal representation (e.g., commonality, particularity, and disturbance), which limit their performance. In order to solve the above issues, in this paper, we propose a novel robust, discriminative and comprehensive dictionary learning (RDCDL) method, in which a robust dictionary is learned from comprehensive training sample diversities generated by extracting and generating facial variations. Especially, to completely represent the commonality, particularity and disturbance, class-shared, class-specific and disturbance dictionary atoms are learned to represent the data from different classes. Discriminative regularizations on the dictionary and the representation coefficients are used to exploit discriminative information, which effectively improves the classification capability of the dictionary. The proposed RDCDL method is extensively evaluated on benchmark face image databases, and it shows superior performance to many state-of-the-art dictionary learning methods for face recognition.

1. Introduction

Inspired by sparse coding mechanism of human vision system, sparse representation represents a signal or an image vector as a sparse linear combination of representation bases which are atoms of the dictionary. Recently sparse representation technology has been successfully used in image restoration [1,2], image classification [3,4,7], and face recognition [5,6], etc. For the success of sparse representation, the dictionary is very important and should effectively represent the encoded signal or image vector [28]. As the dictionary, the analytically designed off-the-shelf bases (e.g., wavelets) might be universal to all types of images, but it will not be effective enough for specific tasks such as face recognition. Instead, many latest methods that learn properly the desired dictionary from the original training data have led to state-of-the-art results in many practical applications, which include image reconstruction [1,8], face recognition [10–12,14,15,21,36], and image classification [8,13,37,49].

Dictionary learning aims to learn the desired dictionary from the training samples. Basically the desired dictionary should well represent or code the given signal. One representative unsupervised dictionary learning model is the KSVD algorithm [16] that learns an over-complete dictionary from a set of image patches. Another unsupervised dictionary learning model is the analysis-synthesis dictionary learning method which learns a pair of dictionaries for image deblurring [47]. According to the relationship between dictionary atoms and class labels, current supervised dictionary learning can be categorized into three main types: class-shared dictionary learning, class-specific dictionary learning and hybrid dictionary learning. For class-shared dictionary learning, each dictionary atom can be used to represent all classes of data. For class-specific dictionary learning, each dictionary atom should be corresponded to only a single class. For hybrid dictionary learning, the hybrid dictionary includes class-shared dictionary and class-specific dictionary.

* Corresponding author at: School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China.
E-mail address: yangm6@mail.sysu.edu.cn (M. Yang).

https://doi.org/10.1016/j.patcog.2018.03.021
0031-3203/

In the first category, a dictionary whose atoms are shared by all classes of data is learned while the discrimination of coding coefficients is exploited [8,10,11,19]. Marial et al. [19] proposed to learn simultaneously discriminative dictionaries with linear classifiers of coding coefficients. Based on KSVD [16], Zhang and Li [10] proposed a dictionary learning method called discriminative KSVD (DKSVD). Following DKSVD [10], Jiang et al. [11] added a label consistent term and proposed so-called label-consistent KSVD (LCKSVD). Recently, Mairal et al. [8] proposed a task-driven dictionary learning framework which minimized different risk functions of the representation coefficients for different tasks. Based on analysis dictionary learning (ADL) [40], Guo et al. [46] proposed discriminative ADL (DADL). Recently, Yang et al. [56] proposed a discriminative model of class-shared analysis and synthesis dictionary pair learning for face recognition. The class-shared dictionary that can represent all classes of data loses the relationship between dictionary atoms and class labels. Thus classifiers based on the class-shared dictionary cannot perform classification based on the class-specific representation residuals, which can weaken the classification capability.

In the second category, class-specific dictionary learning requires that each dictionary atom should correspond to a single class label, so that the class-specific reconstruction error can be used for classification [20–22,26,36]. Wright et al. [5] proposed to use the whole training set to sparsely encode a testing face image, and then classify the testing image by evaluating which class leads to the minimal class-specific reconstruction error. The sparse representation based classification (SRC) [5] framework has shown impressive face recognition results. Inspired by SRC, the class-specific dictionary is widely applied to the design of classifiers. Based on the KSVD [16] model, Mairal et al. [22] introduced a discriminative reconstruction penalty term. Yang et al. [17] and Sprechmann and Sapiro [18] learned a dictionary of sparse representation for each class. In order to encourage the dictionaries of different classes to be independent to each other, Ramirez et al. [20] proposed a model of dictionary learning with structured incoherence (DLSI) which minimized the coherence term of the dictionary to improve the discriminative capability of the dictionary. In action recognition based on images, Castrodad and Sapiro [26] learned a set of action-specific dictionaries with non-negative representation regularization. Yang et al. [21,36] proposed Fisher discrimination dictionary learning (FDDL), where both the representation residual and the representation coefficients achieved discriminative information. Inspired by FDDL, a new analysis and synthesis dictionary pair with Fisher regularized was developed in [57]. Gu et al. [41] proposed a projective class-specific dictionary pair learning algorithm for pattern classification. Although class-specific dictionary learning can achieve good performance, the coherence among the different class-specific sub-dictionaries is inevitable. The number of the dictionaries is usually large.

In the third category, the hybrid dictionary is the dictionary which combines the class-specific dictionary with the class-shared dictionary. Recently, some hybrid dictionary learning methods are proposed. Deng et al. [25] proposed extended sparse representation based classification (ESRC) which constructed an intraclass variation dictionary as a shared dictionary. ESRC achieved promising performance for face recognition with a single sample per person. Wei et al. [39] proposed undersampled face recognition via robust auxiliary dictionary learning. Zhou et al. [13] proposed joint dictionary learning (JDL) where a hybrid dictionary with a Fisher-like regularization on the coding coefficients was learned. Kong et al. [12] proposed dictionary learning with commonality and particularity (COPAR) which learned a hybrid dictionary by introducing an incoherence penalty term to the hybrid dictionary. Shen et al. [27] proposed a hybrid dictionary learning method where the desired dictionary had a hierarchical category structure. Yang et al. [48] proposed a novel dictionary learning method which was analysis-synthesis dictionary learning for universality-particularity representation based classification. Instead of predefining the relationship between dictionary atoms and class labels, Yang et al. [42] proposed a latent dictionary learning (LDL) method to learn a discriminative dictionary and build its relationship to class labels adaptively. However, these hybrid dictionary learning methods cannot well describe the disturbance such as noise, outliers and occlusion. In addition, these methods do not introduce the discriminative information to both the dictionary and the representation coefficients.

Though dictionary learning has achieved promising performance in face recognition, previous dictionary learning methods have some disadvantages. For example, for face recognition, conventional dictionary learning methods cannot well learn a reliable and robust dictionary due to suffering from the small-sample-size problem. Limited number of training samples cannot provide reliable information of face identity and variations so that the learned dictionary may not be robust in the practical application. Some methods about learning an occlusion dictionary [9,23,24] are proposed to recognize the occluded face images and achieve robust performance. However, they may not well handle the general variation in the practical face recognition. The other significant issue is that current dictionary learning do not completely cover the important components of signal representation (e.g., commonality, particularity, and disturbance), which limit their performance. In order to address the above problems, in the paper, we propose a novel robust, discriminative and comprehensive dictionary learning (RDCDL) model.

We propose RDCDL to use the training sample diversities of the same face image to get a robust dictionary. To achieve the robustness, RDCDL learns the dictionary from sample diversities by extracting real face variations and generating virtual face images that convey new possible variations, such as poses, corruption, and occlusion of the face. From original training samples, extracted face variations and virtual training samples, RDCDL learns the dictionary including class-shared dictionary, class-specific dictionary and disturbance dictionary in order to completely represent the practical data (e.g., the data of the different classes has class-shared components, class-specific components and disturbance components such as noise, outliers and occlusion). At the same time, the discriminative regularizations on the dictionary and the representation coefficients have exploited the discriminative information, which effectively improves the discriminative capability of the dictionary.

Although Xu et al. [50] proposed a dictionary learning framework which also used training sample diversities of the same face image and tried to obtain effective representations of face images and a robust dictionary, our proposed RDCDL is quite different from the framework. First, different from the class-shared dictionary learned in [50], we learn a more complete dictionary to represent the commonality, particularity and disturbance of signals. Especially the disturbance dictionary will well represent the disturbance component, with the clean part of signal represented by class-shared dictionary and class-specific dictionary. Second, our proposed model can use the powerful class-specific reconstruction error as the classification criterion which is not used in [50]. Third, apart from the sample diversities simulated by doing with the original training face images only used in [50], the practical face variations are extracted in our paper. What's more, in the proposed RDCDL, the discriminative information is introduced to the dictionary and the representation coefficients.

The rest of this paper is organized as follows. Section 2 briefly introduces related works. Section 3 presents the proposed RDCDL model. Section 4 describes the optimization procedure of RDCDL. Section 5 presents the RDCDL based classification. Section 6 con-

ducts experiments and discusses the RDCDL with deep features. Finally, Section 7 concludes the paper.

2. Brief review of related works

2.1. SRC and ESRC

Wright et al. [5] proposed the sparse representation based classification (SRC) method for robust face recognition. Suppose that there are N classes of subjects. Let $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_N]$ be the set of training samples, where $\boldsymbol{A}_i$ is the subset of the training samples from the i-th class. Denote by $\boldsymbol{y}$ a testing sample. We can sparsely code $\boldsymbol{y}$ over $\boldsymbol{A}$. In this case, the coding coefficient can be got by solving the following equation:

$$\hat{\boldsymbol{\alpha}}=\arg\min_{\boldsymbol{\alpha}}\left\{\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1\right\}, \tag{1}$$

where λ is a scalar constant, $\hat{\boldsymbol{\alpha}}=[\hat{\boldsymbol{\alpha}}_1;\hat{\boldsymbol{\alpha}}_2;\ldots;\hat{\boldsymbol{\alpha}}_N]$. The reconstruction error of each class is represented as follows:

$$e_i=\left\|\boldsymbol{y}-\boldsymbol{A}_i\hat{\boldsymbol{\alpha}}_i\right\|_2, \tag{2}$$

where $\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector associated with the i-th class. SRC utilizes the reconstruction error e_i associated with each class to do classification. The classification is defined as follows:

$$\text{identity}(\boldsymbol{y})=\arg\min_i\{e_i\} \tag{3}$$

Deng et al. [25] proposed ESRC to deal with occlusions by constructing an intra-class variant dictionary $\boldsymbol{D}$ shared by different subjects. In ESRC, a testing sample $\boldsymbol{y}$ can be sparsely coded over $\boldsymbol{A}$ and $\boldsymbol{D}$. In this case, the coding coefficient can be got by solving the following equation:

$$\begin{bmatrix}\hat{\boldsymbol{x}}\\ \hat{\boldsymbol{\beta}}\end{bmatrix}=\arg\min_{\boldsymbol{x},\boldsymbol{\beta}}\left\{\left\|\boldsymbol{y}-[\boldsymbol{A},\boldsymbol{D}]\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{\beta}\end{bmatrix}\right\|_2^2+\lambda\left\|\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{\beta}\end{bmatrix}\right\|_1\right\}, \tag{4}$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are the coding coefficients of $\boldsymbol{y}$ over $\boldsymbol{A}$, $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ are the coding coefficients of $\boldsymbol{y}$ over $\boldsymbol{D}$. The reconstruction error of each class is represented as follows:

$$e_i=\left\|\boldsymbol{y}-[\boldsymbol{A},\boldsymbol{D}]\begin{bmatrix}\delta_i(\hat{\boldsymbol{x}})\\ \hat{\boldsymbol{\beta}}\end{bmatrix}\right\|_2, \tag{5}$$

where $\delta_i(\hat{\boldsymbol{x}})$ is a new vector whose only nonzero entries are the entries in $\hat{\boldsymbol{x}}$ which are associated with the i-th class. The criterion of classification is Eq. (3).

2.2. Hybrid dictionary learning

Recently, hybrid dictionary learning [12,13,25,27], which combines the class-specific dictionary with the class-shared dictionary, becomes popular in the dictionary learning based pattern classification. For instance, Kong et al. [12] proposed dictionary learning with commonality and particularity (COPAR) which learned a hybrid dictionary by introducing an incoherence penalty term to the hybrid dictionary.

Suppose that there are N classes of subjects. Let $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_N]$ is the set of training samples, where $\boldsymbol{A}_j$ $(j=1,\ldots,N)$ is the subset of the training samples from the j-th class and a training sample $\boldsymbol{a}_i$ belongs the j-th class indexed by $i\in\chi_j$. COPAR learned a hybrid dictionary $\boldsymbol{D}=[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N,\boldsymbol{D}_{N+1}]$, where $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$ is the class-specific dictionary, $\boldsymbol{D}_{N+1}$ is the class-shared dictionary. The learned $\boldsymbol{D}$ should well represent every sample $\boldsymbol{a}_i$ as $\boldsymbol{a}_i\approx\boldsymbol{D}\boldsymbol{\theta}_i$, where $\boldsymbol{\theta}_i=[\boldsymbol{\theta}_i^1;\boldsymbol{\theta}_i^2;\ldots;\boldsymbol{\theta}_i^N;\boldsymbol{\theta}_i^{N+1}]$, $\boldsymbol{\theta}_i^j$ $(j=1,\ldots,N)$ is the coding coefficient of $\boldsymbol{a}_i$ over $\boldsymbol{D}_j$, $\boldsymbol{\theta}_i^{N+1}$ is the coding coefficient of $\boldsymbol{a}_i$ over $\boldsymbol{D}_{N+1}$. The incoherence term $\psi(\boldsymbol{D}_k,\boldsymbol{D}_l)=\|\boldsymbol{D}_k^T\boldsymbol{D}_l\|_F^2$ $(k\neq l)$ is introduced to the hybrid dictionary. The COPAR model [12] is written as follows:

$$\min_{\boldsymbol{D}}\sum_{j=1}^{N}\sum_{i\in\chi_j}\left\{\left\|\boldsymbol{a}_i-\boldsymbol{D}\boldsymbol{\theta}_i\right\|_2^2+\lambda\left\|\boldsymbol{\theta}_i\right\|_1+\left\|\boldsymbol{a}_i-\boldsymbol{D}_j\boldsymbol{\theta}_i^j-\boldsymbol{D}_{N+1}\boldsymbol{\theta}_i^{N+1}\right\|_2^2\right\}$$
$$+\eta\sum_{k=1}^{N+1}\sum_{l=1,k\neq l}^{N+1}\psi(\boldsymbol{D}_k,\boldsymbol{D}_l), \tag{6}$$

where λ and η are scalar constants.

The hybrid dictionary learning methods only learn the class-specific dictionary and the class-shared dictionary, which can well represent the commonality and particularity of data. However, the disturbance part of data, such as noise, outliers and occlusion, is neglected in these models.

3. Model of robust, discriminative and comprehensive dictionary learning

In order to improve the performance of previous dictionary learning methods, we propose a new robust, discriminative and comprehensive dictionary learning (RDCDL) model. Suppose that there are N classes of subjects. The RDCDL model learns the comprehensive dictionary $\boldsymbol{D}$ including a class-shared dictionary $\boldsymbol{D}_c$, a class-specific dictionary $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$ and two disturbance dictionaries $\boldsymbol{D}_b$ (i.e., the simulated disturbance dictionary) and $\boldsymbol{D}_p$ (i.e., the real disturbance dictionary). The class-shared dictionary can represent all classes of data. The class-specific dictionary can only represent the particularity of data from a certain class, so that the high-performance class-specific representation residual can be used as the criterion of classification. The disturbance dictionary can represent the disturbance components (e.g., noise, outliers and occlusion) of data. Although the disturbance components have no direct contribution to the final classification, they are very important to the representation of facial images, which has close relation with the final classification. For instance, when the facial occlusion component is represented on the class-shared dictionary and the class-specific dictionary, the representation coefficients associated to these two kinds of dictionaries will be destroyed since they cannot well recover the identity of the query face image. In the ideal case, a clean face image without disturbances is expected to be only represented on the class-shared dictionary and the class-specific dictionary, so that the identity of the query face image can be well recovered based on the representation coefficients. However, in the practical face recognition, there are various disturbances (e.g., noise, lighting changes, expression changes, occlusion, etc.) happened in the query face images. To make the class-shared dictionary and the class-specific dictionary only represent the clean face components, in this paper, two kinds of disturbance dictionaries are introduced to be learned jointly with the other two kinds of dictionaries.

3.1. Extraction of real variation for $\mathrm{D_p}$

Denote by $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_N]\in\Re^{d\times S}$ the set of training samples, where $\boldsymbol{A}_i\in\Re^{d\times S_i}$ $(S=\sum_{i=1}^{N}S_i)$ is the training samples from the i-th class. Denote by $\boldsymbol{G}=[\boldsymbol{G}_1,\boldsymbol{G}_2,\ldots,\boldsymbol{G}_L]\in\Re^{d\times T}$ $(L\leq N)$ the set of generic subject samples, the subjects of $\boldsymbol{G}$ are all different from the subjects of $\boldsymbol{A}$. We use the matrix low rank decomposition [29] method to extract the disturbance components (e.g., noise, outliers, and occlusion) from each class images of $\boldsymbol{G}$. Here, we suppose that the disturbance component only accounts for a small part of the image feature, i.e., the sparse component of the image. We take $\boldsymbol{G}_j\in\Re^{d\times T_j}$ $(T=\sum_{i=1}^{L}T_j)$ as an example, $\boldsymbol{G}_j$ is decomposed as follows:

$$\min_{\boldsymbol{\Lambda}_j,\boldsymbol{E}_j}\text{rank}(\boldsymbol{\Lambda}_j)+\gamma\left\|\boldsymbol{E}_j\right\|_0\ \ s.t.\ \ \boldsymbol{G}_j=\boldsymbol{\Lambda}_j+\boldsymbol{E}_j, \tag{7}$$

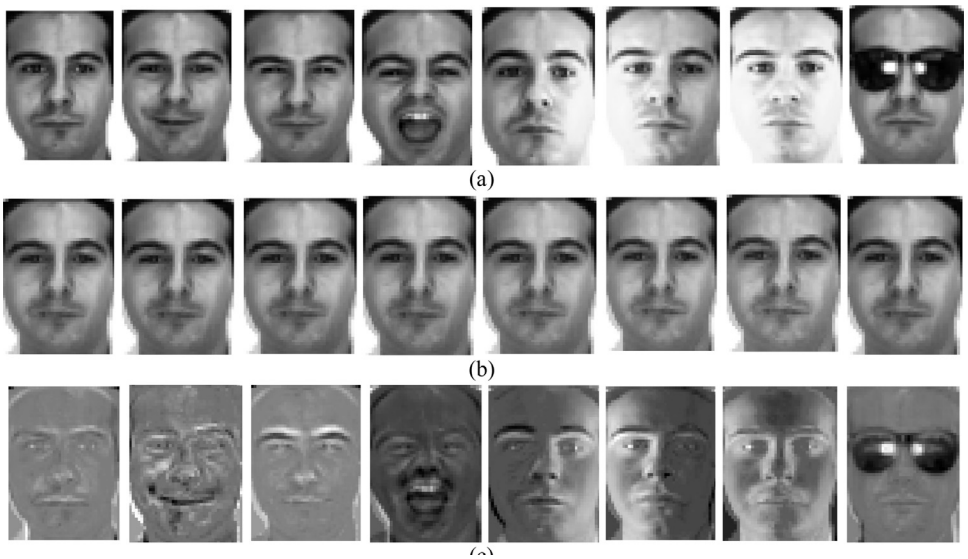

Fig. 1. The images of $\boldsymbol{G}_j$, $\boldsymbol{\Lambda}_j$ and $\boldsymbol{E}_j$. (a) The images of $\boldsymbol{G}_j$. (b) The images of $\boldsymbol{\Lambda}_j$. (c) The images of $\boldsymbol{E}_j$.

Fig. 2. The left two images are the original training samples, the right two images are the alternative training samples by corrupting the left two images using the Salt & Pepper noise.

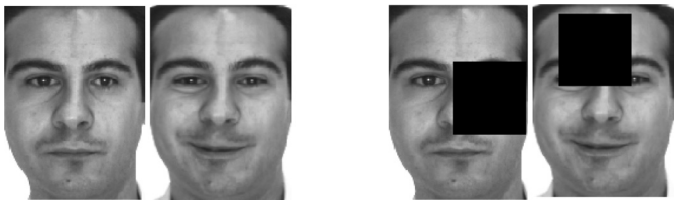

Fig. 3. The left two images are the original training samples, the right two images are the alternative training samples by occluding the left two images using the random square block occlusion.

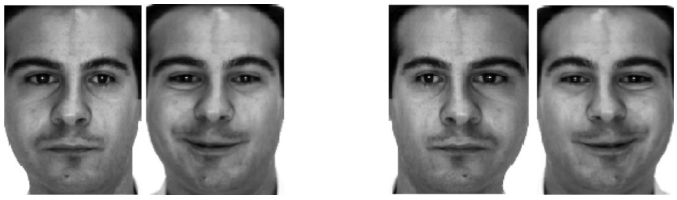

Fig. 4. The left two images are the original training samples, the right two images are the alternative training samples by mirroring the left two images.

where $\gamma > 0$ is a scalar constant that trades off the rank of the solution versus the sparsity of the error, $\boldsymbol{\Lambda}_j$ represents the approximate clear images of $\boldsymbol{G}_j$, while $\boldsymbol{E}_j \in \Re^{d\times T_j}$ $(T = \sum_{j=1}^{L} T_j)$ represents the disturbance components (e.g., noise, outliers, and occlusion) of $\boldsymbol{G}_j$. Denote by $\boldsymbol{E} = [\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_L] \in \Re^{d\times T}$ the set of the disturbance components. Fig. 1 shows the images of $\boldsymbol{G}_j$, $\boldsymbol{\Lambda}_j$ and $\boldsymbol{E}_j$.

3.2. Generation of simulated variation for D_b

Denote by $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N] \in \Re^{d\times S}$ the set of alternative training samples with simulated facial variation. The alternative training samples $\boldsymbol{Z}_i \in \Re^{d\times S_i}$ $(S = \sum_{i=1}^{N} S_i)$ has the same size and structure as the original training samples $\boldsymbol{A}_i$. In order to make the learned dictionary robust to the variations of facial poses and expressions, illuminations and disguises of the same person, we obtain $\boldsymbol{Z}$ by using a special scheme. In this paper, the procedures to generate the alternative training samples are presented as follows:

[1] We take the corrupted images of original training samples as the alternative training samples by using the Matlab function "imnoise" on the original face images. We can obtain the alternative training samples by corrupting the original face images by using salt & pepper noise. For example, $\boldsymbol{a}' = \text{imnoise}(\boldsymbol{a}, \text{'salt \& pepper'}, 0.2)$, where $\boldsymbol{a}$ and $\boldsymbol{a}'$ are the original face image and the alternative training sample, respectively, and 0.2 is the density of noise. Fig. 2 shows the original training samples and the alternative training samples.

[2] We use the original training samples with the random square block occlusion (i.e., the location and size of the block is random) as the alternative training samples. The gray values of the random square block occlusion are set as 0. Fig. 3 shows the original training samples and the alternative training samples.

[3] We use the mirror face images of original training samples as the alternative training samples. For a original training image $\boldsymbol{a}$, its mirror face image is defined as follows:

$$\boldsymbol{a}^*(i, j) = \boldsymbol{a}(i, r - j + 1), (i = 1, \ldots, q; j = 1, \ldots, r), \tag{8}$$

where q and r are the numbers of the rows and columns of the face image matrix, respectively. $\boldsymbol{a}(i, j)$ and $\boldsymbol{a}^*(i, j)$ represent the pixels located in the i-th row and j-th column of $\boldsymbol{a}$ and $\boldsymbol{a}^*$, respectively. Fig. 4 shows the original training samples and the alternative training samples.

3.3. RDCDL model

Here, we denote $\boldsymbol{C} = [\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_N] \in \Re^{K_c\times S}$, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_N] \in \Re^{K\times S}$, $\boldsymbol{B} = [\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_N] \in \Re^{K_b\times S}$, and $\boldsymbol{P} = [\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_L] \in \Re^{K_p\times T}$. For the comprehensive dictionary $\boldsymbol{D} = [\boldsymbol{D}_c, \boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N, \boldsymbol{D}_b, \boldsymbol{D}_p] \in \Re^{d\times(K_c+K+K_b+K_p)}$, we propose the

RDCDL model:

$$J_{(\boldsymbol{D},\boldsymbol{C},\boldsymbol{X},\boldsymbol{B},\boldsymbol{P})} = \arg\min_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{X},\boldsymbol{B},\boldsymbol{P}} \sum_{i=1}^{N} \begin{bmatrix} \left\| \boldsymbol{A}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \sum_{j=1,j\neq i}^{N} \boldsymbol{D}_j\boldsymbol{X}_i^j \right\|_F^2 \\ +\lambda_1 \left\| \boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i \right\|_F^2 \\ +\lambda_2(\|\boldsymbol{C}_i\|_1 + \|\boldsymbol{X}_i\|_1 + \|\boldsymbol{B}_i\|_1) + \lambda_3\phi(\boldsymbol{X}_i) \end{bmatrix} + \sum_{j=1}^{L} \left(\left\| \boldsymbol{E}_j - \boldsymbol{D}_p\boldsymbol{P}_j \right\|_F^2 + \lambda_2 \left\| \boldsymbol{P}_j \right\|_1 \right) + \lambda_4\varphi(\boldsymbol{D}), \quad (9)$$

where λ_1, λ_2, $\lambda_3$and λ_4 are scalar parameters, $\boldsymbol{X}_i^i$ is the coding coefficient matrix of $\boldsymbol{A}_i$ over the dictionary $\boldsymbol{D}_i$, $\boldsymbol{X}_i^j$ is the coding coefficient matrix of $\boldsymbol{A}_i$ over the dictionary $\boldsymbol{D}_j$. In Eq. (9), $\boldsymbol{C}_i \in \Re^{K_c \times S_i}$ and $\boldsymbol{X}_i \in \Re^{K \times S_i}$ $(S = \sum_{i=1}^{N} S_i)$ are the coding coefficient matrices of $\boldsymbol{A}_i$ over the dictionary $\boldsymbol{D}_c \in \Re^{d \times K_c}$ and the dictionary $[\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_N] \in \Re^{d \times K}$, respectively. $\boldsymbol{B}_i \in \Re^{K_b \times S_i}$ $(S = \sum_{i=1}^{N} S_i)$ is the coding coefficient matrix of $\boldsymbol{Z}_i$ over the simulated disturbance dictionary $\boldsymbol{D}_b \in \Re^{d \times K_b}$, while $\boldsymbol{P}_j \in \Re^{K_p \times T_j}$ $(T = \sum_{j=1}^{L} T_j)$ is the coding coefficient matrix of $\boldsymbol{E}_j$ over the real disturbance dictionary $\boldsymbol{D}_p \in \Re^{d \times K_p}$. $\phi(\boldsymbol{X}_i)$ is the representation coefficient discrimination constraint term and $\varphi(\boldsymbol{D})$ is the dictionary discrimination constraint term.

For the i-th class, $\boldsymbol{X}_i = [\boldsymbol{X}_i^1; \boldsymbol{X}_i^2; \ldots; \boldsymbol{X}_i^N] \in \Re^{K \times S_i}$ is the coding coefficient matrix of $\boldsymbol{A}_i$ over the dictionary $[\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N]$, where $\boldsymbol{X}_i^i \in \Re^{K_i \times S_i}$ $(K = \sum_{i=1}^{N} K_i)$ is the coding coefficient matrix of $\boldsymbol{A}_i$ over the dictionary $\boldsymbol{D}_i$. In order to improve the classification capability of the representation coefficient, we require that $\boldsymbol{A}_i$ should be only represented over $\boldsymbol{D}_i$ and not be represented over the other class-special sub-dictionaries, i.e., $\boldsymbol{X}_i^j = 0$ $(i \neq j)$. Therefore the term $\sum_{j\neq i} \boldsymbol{D}_j\boldsymbol{X}_i^j$ will be eliminated.

For the alternative training samples $\boldsymbol{Z}_i$ generated by using ways described in Section 3.2, we require that the representation coefficients of $\boldsymbol{Z}_i$ over $\boldsymbol{D}_c$ and $\boldsymbol{D}_i$ are the same to those of $\boldsymbol{A}_i$ over $\boldsymbol{D}_c$ and $\boldsymbol{D}_i$. For instance, the representation coefficients of $\boldsymbol{Z}_i$ over $\boldsymbol{D}_c$ and $\boldsymbol{D}_i$ and the representation coefficients of $\boldsymbol{A}_i$ over $\boldsymbol{D}_c$ and $\boldsymbol{D}_i$ are $\boldsymbol{C}_i$ and $\boldsymbol{X}_i^i$, respectively. This requirement will protect the representation coefficients of the clean facial components over $\boldsymbol{D}_c$ and $\boldsymbol{D}_i$, with the simulated disturbance components represented by the simulated disturbance dictionary $\boldsymbol{D}_b$.

3.4. Discriminative regularization of $\phi(\mathrm{X_i})$ and $\varphi(\mathrm{D})$

At the same time, we also require that the within-class scatter of the representation coefficients $\boldsymbol{X}_i^i$ should be small, i.e., the representation coefficients of data from the same class over the class-special sub-dictionary should be similar. Thus, the discrimination constraint of $\boldsymbol{X}_i$ can be defined as:

$$\phi(\boldsymbol{X}_i) = \left\| \boldsymbol{X}_i^i - \boldsymbol{M}_i \right\|_F^2, \quad (10)$$

where $\boldsymbol{M}_i$ is the coefficient mean value matrix, each column of $\boldsymbol{M}_i$ is the column mean vector of the representation coefficient matrix $\boldsymbol{X}_i^i$. Because the sparse constraint on $\boldsymbol{X}_i^j$ $(i \neq j)$ results in $\boldsymbol{X}_i^j = 0$, here, we do not show $\boldsymbol{X}_i^j = 0$.

In order to improve the discriminative capability of the dictionary, the coherence among the different class-special sub-dictionaries should be very small, i.e., $\|\boldsymbol{D}_i{}^T\boldsymbol{D}_j\|_F^2$ is small for $i \neq j$, $\|\boldsymbol{D}_b{}^T\boldsymbol{D}_i\|_F^2$ and $\|\boldsymbol{D}_p{}^T\boldsymbol{D}_i\|_F^2$ are also small. Therefore the dictionary discrimination constraint term is designed as follows:

$$\varphi(\boldsymbol{D}) = \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} \left\| \boldsymbol{D}_j{}^T\boldsymbol{D}_i \right\|_F^2 + \sum_{i=1}^{N} \left(\left\| \boldsymbol{D}_b{}^T\boldsymbol{D}_i \right\|_F^2 + \left\| \boldsymbol{D}_p{}^T\boldsymbol{D}_i \right\|_F^2 \right) \quad (11)$$

By incorporating Eqs. (10) and (11) into Eq. (9) and the discrimination representation coefficient constraint $\boldsymbol{X}_i^j = 0$, $\forall i \neq j$, we have the RDCDL model:

$$J_{(\boldsymbol{D},\boldsymbol{C},\boldsymbol{X},\boldsymbol{B},\boldsymbol{P})} = \arg\min_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{X},\boldsymbol{B},\boldsymbol{P}} \sum_{i=1}^{N} \begin{bmatrix} \left\| \boldsymbol{A}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i \right\|_F^2 + \lambda_1 \left\| \boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i \right\|_F^2 \\ +\lambda_2\left(\|\boldsymbol{C}_i\|_1 + \left\|\boldsymbol{X}_i^i\right\|_1 + \|\boldsymbol{B}_i\|_1\right) + \lambda_3 \left\|\boldsymbol{X}_i^i - \boldsymbol{M}_i\right\|_F^2 \\ +\lambda_4\left(\sum_{j=1,j\neq i}^{N} \left\|\boldsymbol{D}_j{}^T\boldsymbol{D}_i\right\|_F^2 + \left(\left\|\boldsymbol{D}_b{}^T\boldsymbol{D}_i\right\|_F^2 + \left\|\boldsymbol{D}_p{}^T\boldsymbol{D}_i\right\|_F^2 \right) \right) \end{bmatrix} + \sum_{j=1}^{L} \left(\left\| \boldsymbol{E}_j - \boldsymbol{D}_p\boldsymbol{P}_j \right\|_F^2 + \lambda_2 \left\| \boldsymbol{P}_j \right\|_1 \right) \quad (12)$$

We also require that l_2-norm of each atom of the dictionary $\boldsymbol{D}$ should be less than or equal to 1 (i.e., $\|\boldsymbol{d}\|_2^2 \leq 1$) to avoid the trivial solution. Although the objective function $\boldsymbol{J}$ in Eq. (12) is not jointly convex to (**D, C, X, B, P**), it is convex with respect to each of $\boldsymbol{D}$ (i.e., $\boldsymbol{D}_c$, $\boldsymbol{D}_i$, $\boldsymbol{D}_b$, $\boldsymbol{D}_p$) and (**C, X, B, P**) when the other is fixed. Thus, Eq. (12) can be solved by alternatively optimizing $\boldsymbol{D}$ and (**C, X, B, P**). The detailed optimization procedures are presented in Section 4.

4. Optimization of RDCDL

We can solve Eq. (12) by alternatively optimizing $\boldsymbol{D}$ and (**C, X, B, P**): Updating (**C, X, B, P**) by fixing $\boldsymbol{D}$; Updating $\boldsymbol{D}$ by fixing (**C, X, B, P**).

4.1. Update C, X, B, P

When $\boldsymbol{D}$ is fixed, the optimization of (**C, X, B, P**) in Eq. (12) is the convex sparse coding problem. To simplify the optimization, we solve **C, X, B** and **P** one by one.

When $\boldsymbol{D}, \boldsymbol{C}, \boldsymbol{B}, \boldsymbol{P}, \boldsymbol{X}_j^j$ $(j = 1, 2, \ldots, N, j \neq i)$ are fixed, we can update $\boldsymbol{X}_i^i$ $(i = 1, 2, \ldots, N)$ atom by atom. The objective function $\boldsymbol{J}$ in Eq. (12) is reduced to:

$$J_{(\boldsymbol{X}_i^i)} = \arg\min_{(\boldsymbol{X}_i^i)} \left\{ Q_1\left(\boldsymbol{X}_i^i\right) + 2\tau \left\| \boldsymbol{X}_i^i \right\|_1 \right\} \quad (13)$$

where $Q_1(\boldsymbol{X}_i^i) = \|\boldsymbol{A}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2 + \lambda_1\|\boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i\|_F^2 + \lambda_3\|\boldsymbol{X}_i^i - \boldsymbol{M}_i\|_F^2$, $\tau = \frac{\lambda_2}{2}$.

Similarly, when $\boldsymbol{D}, \boldsymbol{X}, \boldsymbol{B}, \boldsymbol{P}, \boldsymbol{C}_j (j = 1, 2, \ldots, N, j \neq i)$ are fixed, we can update $\boldsymbol{C}_i$ $(i = 1, 2, \ldots, N)$ atom by atom. When $\boldsymbol{D}, \boldsymbol{C}, \boldsymbol{X}, \boldsymbol{P}, \boldsymbol{B}_j$ $(j = 1, 2, \ldots, N, j \neq i)$ are fixed, we can update $\boldsymbol{B}_i$ $(i = 1, 2, \ldots, N)$ atom by atom. When $\boldsymbol{D}, \boldsymbol{C}, \boldsymbol{X}, \boldsymbol{B}, \boldsymbol{P}_i (i = 1, 2, \ldots, L, i \neq j)$ are fixed, we can update $\boldsymbol{P}_j$ $(j = 1, 2, \ldots, L)$ atom by atom. Thus, the objective function $\boldsymbol{J}$ in Eq. (12) is reduced respectively to:

$$J_{(\boldsymbol{C}_i)} = \arg\min_{(\boldsymbol{C}_i)} \{Q_2(\boldsymbol{C}_i) + 2\tau\|\boldsymbol{C}_i\|_1\} \quad (14)$$

$$J_{(\boldsymbol{B}_i)} = \arg\min_{(\boldsymbol{B}_i)} \{Q_3(\boldsymbol{B}_i) + 2\tau\|\boldsymbol{B}_i\|_1\} \quad (15)$$

$$J_{(\boldsymbol{P}_j)} = \arg\min_{(\boldsymbol{P}_j)} \left\{ Q_4\left(\boldsymbol{P}_j\right) + 2\tau \left\| \boldsymbol{P}_j \right\|_1 \right\}, \quad (16)$$

where $Q_2(\boldsymbol{C}_i) = \|\boldsymbol{A}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2 + \lambda_1\|\boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i\|_F^2$, $Q_3(\boldsymbol{B}_i) = \lambda_1\|\boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i\|_F^2$,$Q_4(\boldsymbol{P}_j) = \|\boldsymbol{E}_j - \boldsymbol{D}_p\boldsymbol{P}_j\|_F^2$. The iterative projection method (IPM) [30] can be used to solve Eqs. (13) – (16). For instance, the IPM algorithm to solve the problem of $J_{(\boldsymbol{X}_i)} = \arg\min_{(\boldsymbol{X}_i)}\{Q(\boldsymbol{X}_i) + 2\tau\|\boldsymbol{X}_i\|_1\}$ is shown in Table 1.

4.2. Update D

When the coding coefficients **C, X, B** and **P** are learned, the optimization to $\boldsymbol{D}$ (i.e., updating $\boldsymbol{D}_i$, $\boldsymbol{D}_c$, $\boldsymbol{D}_b$ and $\boldsymbol{D}_p$ one by one) in Eq. (12) is convex. Therefore in this section, we update $\boldsymbol{D}_i$, $\boldsymbol{D}_c$, $\boldsymbol{D}_b$ and $\boldsymbol{D}_p$ one by one when the other three dictionaries are fixed.

Table 1
The update of representation coefficients in RDCDL.

Algorithm of updating representation coefficients in RDCDL
1. Input: σ, $\tau > 0$.
2. Initialize: $\tilde{\boldsymbol{X}}_i^{(1)} = 0$ and $s = 1$.
3. While convergence or the maximal iteration number is not reached **do**
$s = s + 1$
$\tilde{\boldsymbol{X}}_i^{(s)} = \boldsymbol{\Gamma}_{\tau/\sigma}\left(\tilde{\boldsymbol{X}}_i^{(s-1)} - \frac{1}{2\sigma}\nabla Q(\tilde{\boldsymbol{X}}_i^{(s-1)})\right)$
where $\nabla Q(\tilde{\boldsymbol{X}}_i^{(s-1)})$ is the derivative of $Q(\boldsymbol{X}_i)$ with regard to $\tilde{\boldsymbol{X}}_i^{(s-1)}$, and $\boldsymbol{\Gamma}_{\tau/\sigma}$ is a component-wise soft thresholding operator and defined as: $[\boldsymbol{\Gamma}_{\tau/\sigma}(\boldsymbol{x})]_j = \begin{cases} 0 & |\boldsymbol{x}_j| \le \tau/\sigma \\ \boldsymbol{x}_j - \mathrm{sign}(\boldsymbol{x}_j)\tau/\sigma & \text{otherwise} \end{cases}$
4. Return $\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{X}}_i^{(s)}$.

4.2.1. *Update* $\mathrm{D_i}$

We can update $\boldsymbol{D}_i$ $(i = 1, 2, \ldots, N)$ atom by atom when $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_c$, $\boldsymbol{D}_b$, $\boldsymbol{D}_p$ and all $\boldsymbol{D}_j$ $(j = 1, 2, \ldots, N, j \neq i)$ are fixed. Thus the objective function $\boldsymbol{J}$ in Eq. (12) reduced to:

$$\boldsymbol{D}_i = \arg\min_{\boldsymbol{D}_i}\left[\left\|\boldsymbol{A}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2 + \lambda_1\left\|\boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i\right\|_F^2\right] + \lambda_4\left[\sum_{j=1, j\neq i}^{N}\left\|\boldsymbol{D}_j{}^T\boldsymbol{D}_i\right\|_F^2 + \left(\left\|\boldsymbol{D}_b{}^T\boldsymbol{D}_i\right\|_F^2 + \left\|\boldsymbol{D}_p{}^T\boldsymbol{D}_i\right\|_F^2\right)\right] \quad (17)$$

Let $\bar{\boldsymbol{A}}_i = \boldsymbol{A}_i - \boldsymbol{D}_c\boldsymbol{C}_i$, $\boldsymbol{D}_{-i} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{i-1}, \boldsymbol{D}_{i+1}, \ldots, \boldsymbol{D}_N]$ and $\bar{\boldsymbol{Z}}_i = \boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_b\boldsymbol{B}_i$, then Eq. (17) can be rewritten as:

$$\boldsymbol{D}_i = \arg\min_{\boldsymbol{D}_i}\left[\left\|\bar{\boldsymbol{A}}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2 + \lambda_1\left\|\bar{\boldsymbol{Z}}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2\right] + \lambda_4\left[\sum_{j=1, j\neq i}^{N}\left\|\boldsymbol{D}_j{}^T\boldsymbol{D}_i\right\|_F^2 + \left(\left\|\boldsymbol{D}_b{}^T\boldsymbol{D}_i\right\|_F^2 + \left\|\boldsymbol{D}_p{}^T\boldsymbol{D}_i\right\|_F^2\right)\right] \quad (18)$$

We can update $\boldsymbol{D}_i = [\boldsymbol{d}_i^1, \boldsymbol{d}_i^2, \ldots, \boldsymbol{d}_i^{K_i}]$ atom by atom, where $\boldsymbol{d}_i^k$ $(k = 1, 2, \ldots, K_i)$ is one of atoms in $\boldsymbol{D}_i$. We denote $\boldsymbol{X}_i^i = [\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_{K_i}] \in \Re^{K_i \times S_i}$, where $\boldsymbol{x}_k \in \Re^{1\times S_i}$ is the k-th row of $\boldsymbol{X}_i^i$. Let $\hat{\boldsymbol{A}}_i = \bar{\boldsymbol{A}}_i - \sum_{j\neq k}\boldsymbol{d}_i^j\boldsymbol{x}_j$ and $\hat{\boldsymbol{Z}}_i = \bar{\boldsymbol{Z}}_i - \sum_{j\neq k}\boldsymbol{d}_i^j\boldsymbol{x}_j$, then we can get:

$$\boldsymbol{d}_i^k = \arg\min_{\boldsymbol{d}_i^k}\left\{f(\boldsymbol{d}_i^k) = \left\|\hat{\boldsymbol{A}}_i - \boldsymbol{d}_i^k\boldsymbol{x}_k\right\|_F^2 + \lambda_1\left\|\hat{\boldsymbol{Z}}_i - \boldsymbol{d}_i^k\boldsymbol{x}_k\right\|_F^2 + \lambda_4\left(\left\|\left(\boldsymbol{d}_i^k\right)^T\boldsymbol{D}_{-i}\right\|_F^2 + \left\|\left(\boldsymbol{d}_i^k\right)^T\boldsymbol{D}_b\right\|_F^2 + \left\|\left(\boldsymbol{d}_i^k\right)^T\boldsymbol{D}_p\right\|_F^2\right)\right\} \quad (19)$$

Let $\partial f(\boldsymbol{d}_i^k)/\partial\boldsymbol{d}_i^k = 0$, then we can obtain the updated $\boldsymbol{d}_i^k$ as follows:

$$\boldsymbol{d}_i^k = \left[(1+\lambda_1)\|\boldsymbol{x}_k\|_2^2\,\boldsymbol{I} + \lambda_4\left(\boldsymbol{D}_{-i}\boldsymbol{D}_{-i}{}^T + \boldsymbol{D}_b\boldsymbol{D}_b{}^T + \boldsymbol{D}_p\boldsymbol{D}_p{}^T\right)\right]^{-1}\left(\hat{\boldsymbol{A}}_i + \lambda_1\hat{\boldsymbol{Z}}_i\right)\boldsymbol{x}_k{}^T \quad (20)$$

As an atom of dictionary, $\boldsymbol{d}_i^k$ should be unitized, i.e. $\hat{\boldsymbol{d}}_i^k = \boldsymbol{d}_i^k/\|\boldsymbol{d}_i^k\|_2$, the corresponding coefficient should be $\hat{\boldsymbol{x}}_k = \|\boldsymbol{d}_i^k\|_2\boldsymbol{x}_k$. When all $\boldsymbol{d}_i^k$ $(k = 1, 2, \ldots, K_i)$ are updated, $\boldsymbol{D}_i$ is learned. With the similar optimization strategy, all class-specific sub-dictionary $\boldsymbol{D}_i$ $(i = 1, 2, \cdots, N)$ are updated.

4.2.2. *Update* $\mathrm{D_c}$

We can update $\boldsymbol{D}_c$ atom by atom when $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_b$, $\boldsymbol{D}_p$ and $[\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N]$ are fixed. Thus the objective function $\boldsymbol{J}$ in Eq. (12) is reduced to:

$$\boldsymbol{D}_c = \arg\min_{\boldsymbol{D}_c}\sum_{i=1}^{N}\left[\left\|\boldsymbol{A}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_F^2 + \lambda_1\left\|\boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i\right\|_F^2\right] \quad (21)$$

Let $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_N]$, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N]$, $\boldsymbol{B} = [\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_N]$, $\boldsymbol{W} = [\boldsymbol{D}_1\boldsymbol{X}_1^1, \boldsymbol{D}_2\boldsymbol{X}_2^2, \ldots, \boldsymbol{D}_N\boldsymbol{X}_N^N]$, $\bar{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{W}$, $\bar{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{W} - \boldsymbol{D}_b\boldsymbol{B}$ and $\boldsymbol{C} = [\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_N]$, then Eq. (21) can be rewritten as:

$$\boldsymbol{D}_c = \arg\min_{\boldsymbol{D}_c}\left\|\bar{\boldsymbol{A}} - \boldsymbol{D}_c\boldsymbol{C}\right\|_F^2 + \lambda_1\left\|\bar{\boldsymbol{Z}} - \boldsymbol{D}_c\boldsymbol{C}\right\|_F^2 \quad (22)$$

As well, we can update $\boldsymbol{D}_c = [\boldsymbol{d}_c^1, \boldsymbol{d}_c^2, \ldots, \boldsymbol{d}_c^{K_c}]$ atom by atom, where $\boldsymbol{d}_c^k$ $(k = 1, 2, \ldots, K_c)$ is one of atoms in $\boldsymbol{D}_c$. We denote $\boldsymbol{C} = [\boldsymbol{t}_1; \boldsymbol{t}_2; \ldots; \boldsymbol{t}_{K_c}] \in \Re^{K_c\times S}$, where $\boldsymbol{t}_k \in \Re^{1\times S}$ is the k-th row of $\boldsymbol{C}$. Let $\hat{\boldsymbol{A}} = \bar{\boldsymbol{A}} - \sum_{j\neq k}\boldsymbol{d}_c^j\boldsymbol{t}_j$ and $\hat{\boldsymbol{Z}} = \bar{\boldsymbol{Z}} - \sum_{j\neq k}\boldsymbol{d}_c^j\boldsymbol{t}_j$, then we can get:

$$\boldsymbol{d}_c^k = \arg\min_{\boldsymbol{d}_c^k}\left\{g(\boldsymbol{d}_c^k) = \left\|\hat{\boldsymbol{A}} - \boldsymbol{d}_c^k\boldsymbol{t}_k\right\|_F^2 + \lambda_1\left\|\hat{\boldsymbol{Z}} - \boldsymbol{d}_c^k\boldsymbol{t}_k\right\|_F^2\right\} \quad (23)$$

Let $\partial g(\boldsymbol{d}_c^k)/\partial\boldsymbol{d}_c^k = 0$, then we can obtain the updated $\boldsymbol{d}_c^k$ as follows:

$$\boldsymbol{d}_c^k = \left[(1+\lambda_1)\|\boldsymbol{t}_k\|_2^2\boldsymbol{I}\right]^{-1}\left(\hat{\boldsymbol{A}} + \lambda_1\hat{\boldsymbol{Z}}\right)\boldsymbol{t}_k{}^T \quad (24)$$

The unitization of $\boldsymbol{d}_c^k$ is $\hat{\boldsymbol{d}}_c^k = \boldsymbol{d}_c^k/\|\boldsymbol{d}_c^k\|_2$ with the corresponding coefficient $\hat{\boldsymbol{t}}_k = \|\boldsymbol{d}_c^k\|_2\boldsymbol{t}_k$. After all dictionary atoms in $\boldsymbol{D}_c$ are updated, the whole dictionary $\boldsymbol{D}_c$ is learned.

4.2.3. *Update* $\mathrm{D_b}$

We can update $\boldsymbol{D}_b$ atom by atom when $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_c$, $\boldsymbol{D}_p$ and $[\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N]$ are fixed. Thus the objective function $\boldsymbol{J}$ in Eq. (12) is reduced to:

$$\boldsymbol{D}_b = \arg\min_{\boldsymbol{D}_b}\sum_{i=1}^{N}\lambda_1\left\|\boldsymbol{Z}_i - \boldsymbol{D}_c\boldsymbol{C}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i - \boldsymbol{D}_b\boldsymbol{B}_i\right\|_F^2 + \lambda_4\sum_{i=1}^{N}\left\|\boldsymbol{D}_b{}^T\boldsymbol{D}_i\right\|_F^2 \quad (25)$$

Let $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N]$, $\tilde{\boldsymbol{D}} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N]$, $\boldsymbol{B} = [\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_N]$, $\boldsymbol{W} = [\boldsymbol{D}_1\boldsymbol{X}_1^1, \boldsymbol{D}_2\boldsymbol{X}_2^2, \ldots, \boldsymbol{D}_N\boldsymbol{X}_N^N]$, $\boldsymbol{C} = [\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_N]$ and $\tilde{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{W} - \boldsymbol{D}_c\boldsymbol{C}$ (note $\tilde{\boldsymbol{Z}}$ here is different from $\bar{\boldsymbol{Z}}$ in the update of $\boldsymbol{D}_c$), then Eq. (25) can be rewritten as:

$$\boldsymbol{D}_b = \arg\min_{\boldsymbol{D}_b}\lambda_1\left\|\tilde{\boldsymbol{Z}} - \boldsymbol{D}_b\boldsymbol{B}\right\|_F^2 + \lambda_4\left\|\boldsymbol{D}_b{}^T\tilde{\boldsymbol{D}}\right\|_F^2 \quad (26)$$

Similarly, we can update $\boldsymbol{D}_b = [\boldsymbol{d}_b^1, \boldsymbol{d}_b^2, \ldots, \boldsymbol{d}_b^{K_b}]$ atom by atom, $\boldsymbol{d}_b^k$ $(k = 1, 2, \ldots, K_b)$ is one of atoms in $\boldsymbol{D}_b$. We denote $\boldsymbol{B} = [\boldsymbol{u}_1; \boldsymbol{u}_2; \ldots; \boldsymbol{u}_{K_b}] \in \Re^{K_b\times S}$, where $\boldsymbol{u}_k \in \Re^{1\times S}$ is the k-th row of $\boldsymbol{B}$. Let $\hat{\boldsymbol{Z}} = \tilde{\boldsymbol{Z}} - \sum_{j\neq k}\boldsymbol{d}_b^j\boldsymbol{u}_j$, then we can get:

$$\boldsymbol{d}_b^k = \arg\min_{\boldsymbol{d}_b^k}\left\{h(\boldsymbol{d}_b^k) = \lambda_1\left\|\hat{\boldsymbol{Z}} - \boldsymbol{d}_b^k\boldsymbol{u}_k\right\|_F^2 + \lambda_4\left\|\left(\boldsymbol{d}_b^k\right)^T\tilde{\boldsymbol{D}}\right\|_F^2\right\} \quad (27)$$

Let $\partial h(\boldsymbol{d}_b^k)/\partial\boldsymbol{d}_b^k = 0$, then the updated $\boldsymbol{d}_b^k$ is described as follows:

$$\boldsymbol{d}_b^k = \left[\lambda_1\|\boldsymbol{u}_k\|_2^2\boldsymbol{I} + \lambda_4\left(\tilde{\boldsymbol{D}}\tilde{\boldsymbol{D}}^T\right)\right]^{-1}\lambda_1\hat{\boldsymbol{Z}}\boldsymbol{u}_k{}^T \quad (28)$$

We can unitize $\boldsymbol{d}_b^k$ to get $\hat{\boldsymbol{d}}_b^k = \boldsymbol{d}_b^k/\|\boldsymbol{d}_b^k\|_2$ with the corresponding coefficient $\hat{\boldsymbol{u}}_k = \|\boldsymbol{d}_b^k\|_2\boldsymbol{u}_k$. After all the dictionary atoms in $\boldsymbol{D}_b$ are updated, the whole dictionary $\boldsymbol{D}_b$ is learned.

Table 2
Robust, discriminative and comprehensive dictionary learning algorithm.

Robust, discriminative and comprehensive dictionary learning
1. Initialize $\boldsymbol{D}=[\boldsymbol{D}_c,\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N,\boldsymbol{D}_b,\boldsymbol{D}_p]$. We use $\boldsymbol{A}_i$ $(i=1,2,\ldots,N)$, $\boldsymbol{Z}-\boldsymbol{A}$ and $\boldsymbol{E}$ as atoms of $\boldsymbol{D}_i$, $\boldsymbol{D}_b$ and $\boldsymbol{D}_p$, respectively. We use PCA to initialize the atoms of $\boldsymbol{D}_c$ by using the set of training samples $\boldsymbol{A}$.
2. Update the representation coefficient *C, X, B, P*. Fix $\boldsymbol{D},\boldsymbol{C},\boldsymbol{B},\boldsymbol{P},\boldsymbol{X}_j^j$ $(j=1,2,\ldots,N,j\neq i)$ and update $\boldsymbol{X}_i^i$ $(i=1,2,\ldots,N)$ atom by atom. Fix $\boldsymbol{D}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{C}_j$ $(j=1,2,\ldots,N,j\neq i)$ and update $\boldsymbol{C}_i$ $(i=1,2,\ldots,N)$ atom by atom. Fix $\boldsymbol{D}$, $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{P}$, $\boldsymbol{B}_j$ $(j=1,2,\ldots,N,j\neq i)$ and update $\boldsymbol{B}_i$ $(i=1,2,\ldots,N)$ atom by atom. Fix $\boldsymbol{D}$, $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}_i$ $(i=1,2,\ldots,L,i\neq j)$ and update $\boldsymbol{P}_j$ $(j=1,2,\ldots,L)$ atom by atom.
3. Update the dictionary $\boldsymbol{D}=[\boldsymbol{D}_c,\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N,\boldsymbol{D}_b,\boldsymbol{D}_p]$. Fix $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_c$, $\boldsymbol{D}_b$, $\boldsymbol{D}_p$ and all $\boldsymbol{D}_j$ $(j=1,2,\ldots,N,j\neq i)$ and update $\boldsymbol{D}_i$ $(i=1,2,\ldots,N)$ atom by atom. Fix $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_b$, $\boldsymbol{D}_p$ and $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$ and update $\boldsymbol{D}_c$ atom by atom. Fix $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_c$, $\boldsymbol{D}_p$ and $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$ and update $\boldsymbol{D}_b$ atom by atom. Fix $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_c$, $\boldsymbol{D}_b$ and $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$ and update $\boldsymbol{D}_p$ atom by atom.
4. Output. Return to step 2 until the values of $J_{(\boldsymbol{D},\boldsymbol{C},\boldsymbol{X},\boldsymbol{B},\boldsymbol{P})}$ between two adjacent iterations are closed enough, or the maximum number of iterations is reached. Output $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$ and $\boldsymbol{D}$.

4.2.4. Update $\mathrm{D_p}$

We can update $\boldsymbol{D}_p$ atom by atom when $\boldsymbol{C}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{D}_c$, $\boldsymbol{D}_b$ and $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$ are fixed. Thus, the objective function $\boldsymbol{J}$ in Eq. (12) is reduced to:

$$\boldsymbol{D}_p=\arg\min_{\boldsymbol{D}_p}\sum_{j=1}^{L}\left\|\boldsymbol{E}_j-\boldsymbol{D}_p\boldsymbol{P}_j\right\|_F^2+\lambda_4\sum_{i=1}^{N}\left\|\boldsymbol{D}_p{}^T\boldsymbol{D}_i\right\|_F^2 \tag{29}$$

Let $\boldsymbol{E}=[\boldsymbol{E}_1,\boldsymbol{E}_2,\ldots,\boldsymbol{E}_L]$, $\tilde{\boldsymbol{D}}=[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$, $\boldsymbol{P}=[\boldsymbol{P}_1,\boldsymbol{P}_2,\ldots,\boldsymbol{P}_L]$, then Eq. (29) can be rewritten as:

$$\boldsymbol{D}_p=\arg\min_{\boldsymbol{D}_p}\|\boldsymbol{E}-\boldsymbol{D}_p\boldsymbol{P}\|_F^2+\lambda_4\left\|\boldsymbol{D}_p{}^T\tilde{\boldsymbol{D}}\right\|_F^2 \tag{30}$$

As well, we can update $\boldsymbol{D}_p=[\boldsymbol{d}_p^1,\boldsymbol{d}_p^2,\ldots,\boldsymbol{d}_p^{K_p}]$ atom by atom, $\boldsymbol{d}_p^k$ $(k=1,2,\ldots,K_p)$ is one of atoms in $\boldsymbol{D}_p$. We denote $\boldsymbol{P}=[\boldsymbol{v}_1;\boldsymbol{v}_2;\ldots;\boldsymbol{v}_{K_p}]\in\Re^{K_p\times T}$, where $\boldsymbol{v}_k\in\Re^{1\times T}$ is the k-th row of $\boldsymbol{P}$. Let $\hat{\boldsymbol{E}}=\boldsymbol{E}-\sum_{j\neq k}\boldsymbol{d}_p^j\boldsymbol{v}_j$, then we can get:

$$\boldsymbol{d}_p^k=\arg\min_{\boldsymbol{d}_p^k}\left\{\rho\left(\boldsymbol{d}_p^k\right)=\left\|\hat{\boldsymbol{E}}-\boldsymbol{d}_p^k\boldsymbol{v}_k\right\|_F^2+\lambda_4\left\|\left(\boldsymbol{d}_p^k\right)^T\tilde{\boldsymbol{D}}\right\|_F^2\right\} \tag{31}$$

Let $\partial\rho(\boldsymbol{d}_p^k)/\partial\boldsymbol{d}_p^k=0$, then the updated $\boldsymbol{d}_p^k$ can be derived as follows:

$$\boldsymbol{d}_p^k=\left[\|\boldsymbol{v}_k\|_2^2\boldsymbol{I}+\lambda_4\left(\tilde{\boldsymbol{D}}\tilde{\boldsymbol{D}}^T\right)\right]^{-1}\hat{\boldsymbol{E}}\boldsymbol{v}_k{}^T \tag{32}$$

The unitization of $\boldsymbol{d}_p^k$ is $\hat{\boldsymbol{d}}_p^k=\boldsymbol{d}_p^k/\|\boldsymbol{d}_p^k\|_2$ with the corresponding coefficient $\hat{\boldsymbol{v}}_k=\|\boldsymbol{d}_p^k\|_2\boldsymbol{v}_k$. After all the dictionary atoms in $\boldsymbol{D}_p$ are updated, the whole dictionary $\boldsymbol{D}_p$ is learned.

4.3. Algorithm of RDCDL

The algorithm of RDCDL is summarized in Table 2. Fig. 5 shows the flowchart of RDCDL. From Fig. 5, we can observe that there are three inputted data, such as labeled training data, real variation and samples with simulated variation. With the proposed robust, discriminative, and comprehensive dictionary learning, the class-shared dictionary $\boldsymbol{D}_c$, the class-specific dictionary $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$, the simulated disturbance $\boldsymbol{D}_b$ and the real disturbance $\boldsymbol{D}_p$ are learned. The testing sample is sparsely represented on the comprehensive dictionary of $\boldsymbol{D}_c$, $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$, $\boldsymbol{D}_b$ and $\boldsymbol{D}_p$, and then classified to the class with minimal reconstruction error.

In the experiments, the number of iterations is set as 10. Since the objective function of RDCDL is lower bounded and can decrease in both updating dictionaries and updating coding coefficients, the algorithm of RDCDL can converge. The objective function value versus the iteration number on different face databases is shown in Fig. 6.

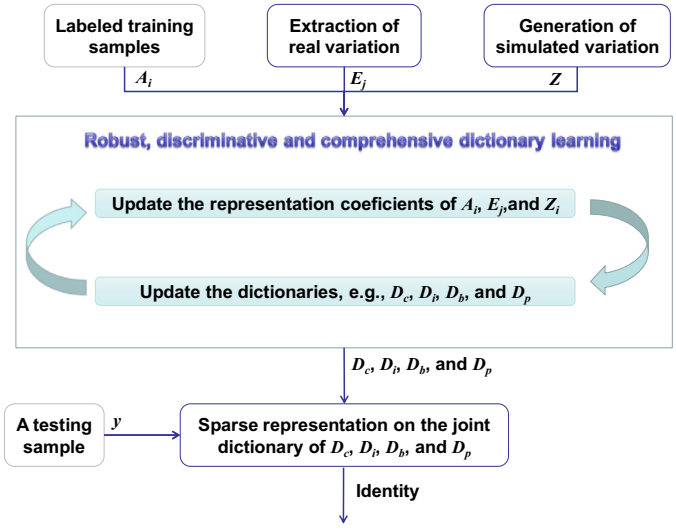

Fig. 5. The flowchart of the proposed RDCDL.

In order to further understand the learned dictionaries, Fig. 7 shows some atoms of class-specific dictionary, atoms of class-shared dictionary, some atoms of the simulated disturbance dictionary and some atoms of the real disturbance dictionary on the AR database with 55×40 face images.

4.4. Time complexity

In the proposed RDCDL algorithm, the time complexity of updating the coding coefficients $\boldsymbol{X}$ is approximately $\sum_i S_i O(d^2K_i{}^{\varepsilon})$ [44], where $\varepsilon\geq1.2$ is a constant, d is the feature dimension, K_i is the number of dictionary atoms in $\boldsymbol{D}_i$, S_i is the number of training samples from the i-th class. $K=\sum_i K_i$ is the number of dictionary atoms in $[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$. $S=\sum_i S_i$ is the total training samples. The time complexity of updating the coding coefficients $\boldsymbol{B}$ is approximately $\sum_i S_i O(d^2K_b{}^{\varepsilon})$, where K_b is the number of dictionary atoms in $\boldsymbol{D}_b$. The time complexity of updating the coding coefficients $\boldsymbol{C}$ is approximately $\sum_i S_i O(d^2K_c{}^{\varepsilon})$, where K_c is the number of dictionary atoms in $\boldsymbol{D}_c$. The time complexity of updating the coding coefficients $\boldsymbol{P}$ is approximately $\sum_j T_j O(d^2K_p{}^{\varepsilon})$, where T_j is the number of disturbance components in $\boldsymbol{E}_j$, $T=\sum_j T_j$ is the number of disturbance components in $\boldsymbol{E}$, K_p is the number of dictionary atoms in $\boldsymbol{D}_p$.

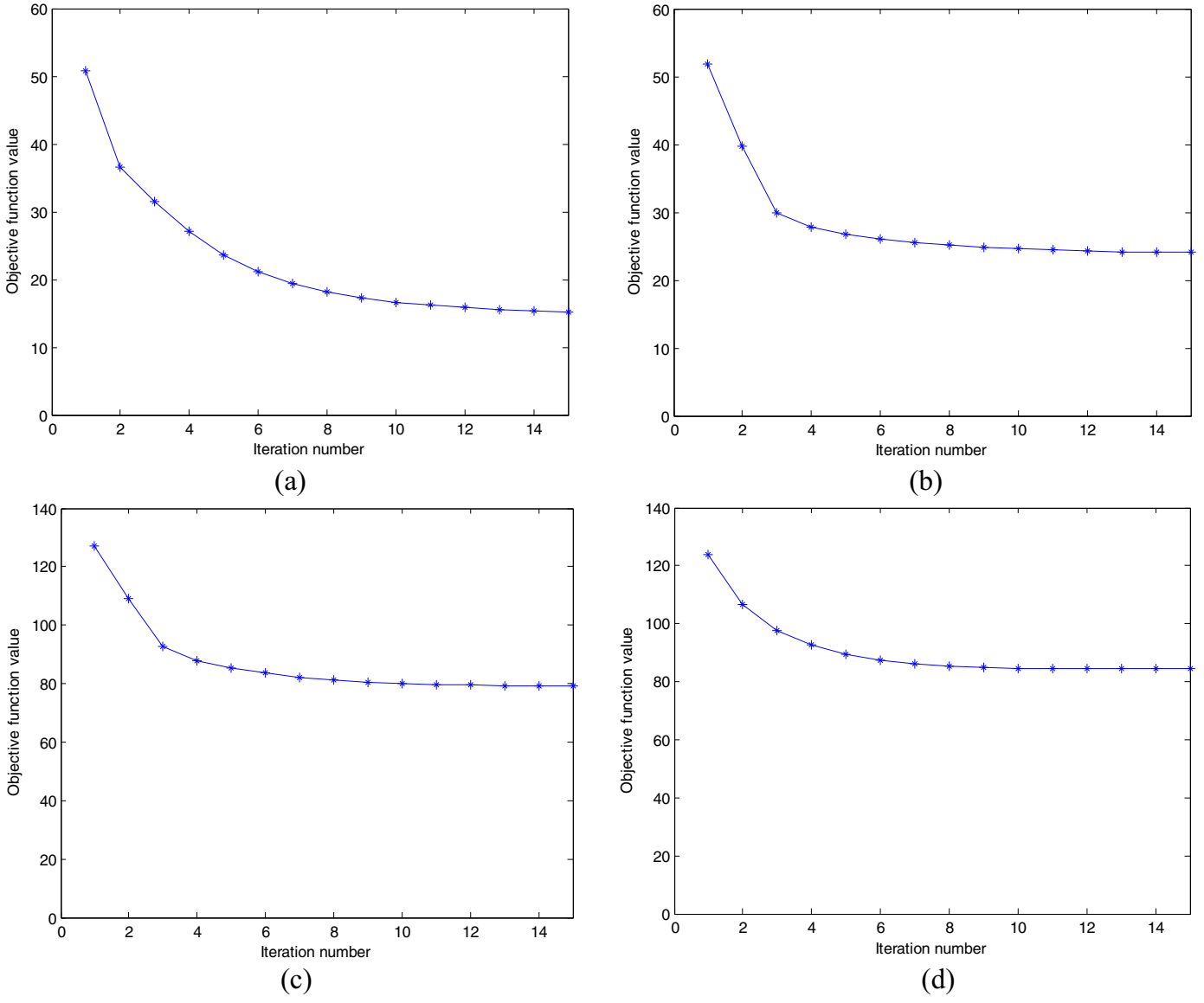

Fig. 6. The objective function value versus the iteration number on different face databases. (a) Extended Yale B. (b) AR. (c) Multi-PIE. (d) LFW.

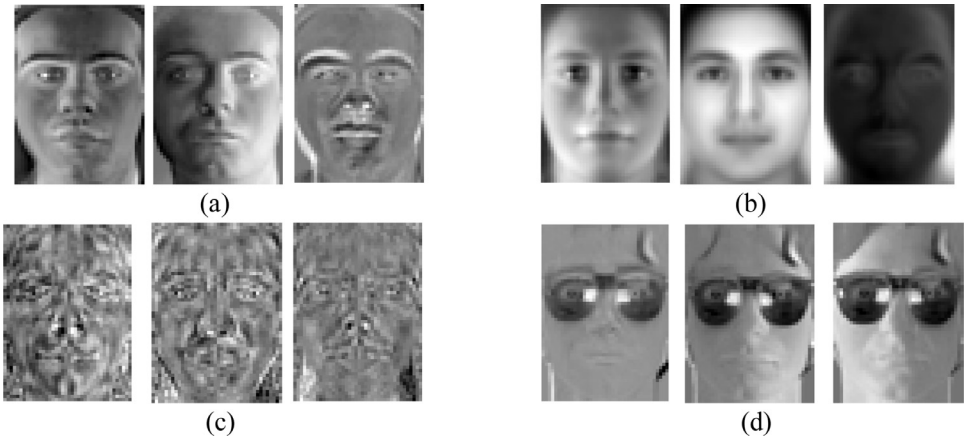

Fig. 7. The learned dictionaries on the AR database. (a) Some atoms of class-specific dictionary. (b) Atoms of class-shared dictionary. (c) Some atoms of the simulated disturbance dictionary $\boldsymbol{D}_b$. (d) Some atoms of the real disturbance dictionary $\boldsymbol{D}_p$.

The time complexity of updating dictionary atoms in $[\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N]$ (i.e., Eq. (20)) is $O(d^2K_b + d^2K_p) + 2O(d^2K) + KO(md^2)$, where $\boldsymbol{D}_b\boldsymbol{D}_b^T + \boldsymbol{D}_p\boldsymbol{D}_p^T$ is computed only once, $\boldsymbol{D}_{-i}\boldsymbol{D}_{-i}^T$ needs compute at most twice, and m is the number of iterations in conjugate gradient method [45] to solve the unknown dictionary atom instead of computing the inverse of a matrix. The time complexity of updating dictionary atoms in $\boldsymbol{D}_c$ (i.e., Eq. (24)) is $K_cO(md^2)$. The time complexity of updating dictionary atoms in $\boldsymbol{D}_b$ (i.e., Eq. (28)) is $K_bO(md^2) + O(d^2K)$, where $\tilde{\boldsymbol{D}}\tilde{\boldsymbol{D}}^T$ is computed once. The time complexity of updating dictionary atoms in $\boldsymbol{D}_p$ (i.e., Eq. (32)) is $K_pO(md^2)$. In total, the time complexity of updating all dictionary atoms is $O(d^2K_b + d^2K_p + 3d^2K) + (K + K_c + K_b + K_p)O(md^2)$.

5. The classification scheme

Once the dictionary $\boldsymbol{D}$ (i.e., the comprehensive dictionary of $\boldsymbol{D}_c$, $[\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N]$, $\boldsymbol{D}_b$ and $\boldsymbol{D}_p$) is learned, it can be used to effectively represent a testing sample $\boldsymbol{y}$ and recognize the identity of $\boldsymbol{y}$. According to the dictionary $\boldsymbol{D}$, different information can be used to perform the classification task. In SRC [5], the original training samples are utilized as a dictionary to represent the testing sample, and the reconstruction error associated with each class is used for classification. In ESRC [25], the original training samples and the intra-class variation dictionary are combined into the hybrid dictionary to represent the testing sample, and the reconstruction error associated with each class is used for classification. In COPAR [12], the class-specific dictionary and the class-shared dictionary are combined into the hybrid dictionary to represent the testing sample, and the reconstruction error associated with each class is used for classification. The classification methods of [5, 12, 25] achieve promising performance in face recognition, so the proposed RDCDL also use the similar classification method.

After the comprehensive dictionary $\boldsymbol{D} = [\boldsymbol{D}_c, \boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N, \boldsymbol{D}_b, \boldsymbol{D}_p]$ is got, we can code a testing sample $\boldsymbol{y}$ over the dictionary $\boldsymbol{D}$. In this case, the coding coefficient can be got by solving:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{y} - [\boldsymbol{D}_c, \boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_N, \boldsymbol{D}_b, \boldsymbol{D}_p] [\boldsymbol{\alpha}_c; \boldsymbol{\alpha}_1; \ldots; \boldsymbol{\alpha}_N; \boldsymbol{\alpha}_b; \boldsymbol{\alpha}_p]\|_2^2 + \lambda\|[\boldsymbol{\alpha}_c; \boldsymbol{\alpha}_1; \ldots; \boldsymbol{\alpha}_N; \boldsymbol{\alpha}_b; \boldsymbol{\alpha}_p]\|_1\}, \quad (33)$$

where λ is a constant. Denote by $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_c; \hat{\boldsymbol{\alpha}}_1; \ldots; \hat{\boldsymbol{\alpha}}_N; \hat{\boldsymbol{\alpha}}_b; \hat{\boldsymbol{\alpha}}_p]$. The reconstruction error of each class is represented as:

$$e_i = \left\|\boldsymbol{y} - \boldsymbol{D}_c\hat{\boldsymbol{\alpha}}_c - \boldsymbol{D}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{D}_b\hat{\boldsymbol{\alpha}}_b - \boldsymbol{D}_p\hat{\boldsymbol{\alpha}}_p\right\|_2, \quad (34)$$

where $\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector associated with the i-th class. The classification is defined as:

$$\text{identity}(\boldsymbol{y}) = \arg\min_i \{e_i\} \quad (35)$$

6. Experimental results and discussion

In order to well show the advantage of RDCDL, we compare it with SVM, SRC [5], CRC [43], DKSVD [10], LCKSVD [11], COPAR [12], FDDL [21], DLSI [20], LDL [42] and ESRC [25] algorithms by experiments on the Extended Yale B [31,32], AR [33], Multi-PIE [34], FRGC [38] and the aligned labeled face in the wild (LFWa) [35].

6.1. Experimental setting

In this section, we give the experimental details. The programming environment is MATLAB R2013a, 3.40 GHz CPU and 8 G RAM. As shown in Eq. (12), there are four parameters (i.e., λ_1, λ_2, λ_3, λ_4) to be determined. λ_1 controls the importance of simulated sample representation, λ_2 determines the sparsity of the representation coefficients, with the discrimination regularized by λ_3 and λ_4. Therefore, we set the values of (λ_1 and λ_2) and (λ_3 and λ_4) alternatively. The cross-validation method is used to select the values of λ_1, λ_2, λ_3, λ_4. We select λ_1, λ_4 from a small set {0.01, 0.001, 0.0001} and λ_2, λ_3 from a small set {0.1, 0.05, 0.01, 0.005, 0.001}. We evaluate the parameter setting on the LFW database (the experiment setting is given in Section 6.6, the dimension of histograms is 500). With fixed $\lambda_1 = 0.001$ and $\lambda_2 = 0.05$, Fig. 8(a) shows the recognition rates versus λ_3 and λ_4. From Fig. 8(a), it can be seen that the recognition rate is not sensitive to the values of λ_3 and λ_4, and $\lambda_3 = 0.05$ and $\lambda_4 = 0.0001$ can achieve good performance. With fixed $\lambda_3 = 0.05$ and $\lambda_4 = 0.0001$, Fig. 8(b) shows the recognition rates versus λ_1 and λ_2. From Fig. 8(b), it can be seen that the recognition rate is not sensitive to the values of λ_1 and λ_2, and $\lambda_1 = 0.001$ and $\lambda_2 = 0.05$ can achieve good performance. In the experiment, for convenience, if no specific instruction is given, the parameters are fixed as $\lambda_1 = 0.001$, $\lambda_2 = 0.05$, $\lambda_3 = 0.05$, $\lambda_4 = 0.0001$ in all the experiments. With the learned dictionary, the parameter of sparse representation is set as λ=0.001.

Table 3
The recognition rates (%) and computing time for training dictionaries and classifying a testing sample on the Extended Yale B database.

Algorithm	100	130	150	TrT (s)	TtT (s)
SVM	30.2	31.1	31.3	–	7.8e–5
SRC[5]	81.7	84.6	85.1	–	5.9e–3
CRC[43]	79.4	84.2	83.6	–	3.4e–4
DKSVD[10]	71.5	71.8	72.9	13.0	3.9e–3
LCKSVD[11]	69.8	70.9	71.2	6.7	9.6e–4
COPAR[12]	81.6	86.3	85.8	7.5	3.5e–4
FDDL[21]	82.8	85.2	85.2	16.4	9.1e–3
DLSI[20]	83.4	85.2	85.9	6.6	4.3e–3
LDL[42]	83.8	84.9	84.7	2.7	3.1e–3
ESRC[25]	85.9	87.9	88.2	–	6.1e–3
RDCDL	**91.6**	**91.6**	**92.4**	4.7	7.1e–3

6.2. Experimental results on the Extended Yale b database

The Extended Yale B database consists of 2414 frontal face images from 38 individuals (about 64 images per subject) captured under various laboratory controlled lighting conditions. Fig. 9 shows images of one person from the Extended Yale B face database. In the experiment, the size of the original face images is 96×84, we select the former 32 subjects and the former 5 images per subject from subset 1 for training and the same 32 subjects from subset 3 and subset 4 for testing. Because there are very drastic illumination changes in the face images of subset 3 and subset 4, the alternative training samples are produced by occluding the original training samples using the random square block occlusion, whose level is 0.2. In this test, the parameters of sparse constraint and discriminative representation are set as $\lambda_2 = 0.005$ and $\lambda_3 = 0.1$. In order to construct the set of disturbance components, the remainder 6 subjects from subset 5 are selected. The disturbance components are computed by Eq. (7). We evaluate the compared methods by reducing the feature dimension of images to 100, 130 and 150 via PCA. The number of dictionary atoms is the same as the number of original training samples.

Table 3 shows the recognition results of the proposed algorithm and ten compared algorithms. In Table 3, "TrT" and "TtT" mean the computing time of training dictionaries and classifying a testing sample when the feature dimension of images is 150. From Table 3, it can be seen that RDCDL achieves higher recognition rates than the compared algorithms. ESRC achieves the second best recognition rates. The recognition rate of RDCDL is 5.7%, 3.7% and 4.2%

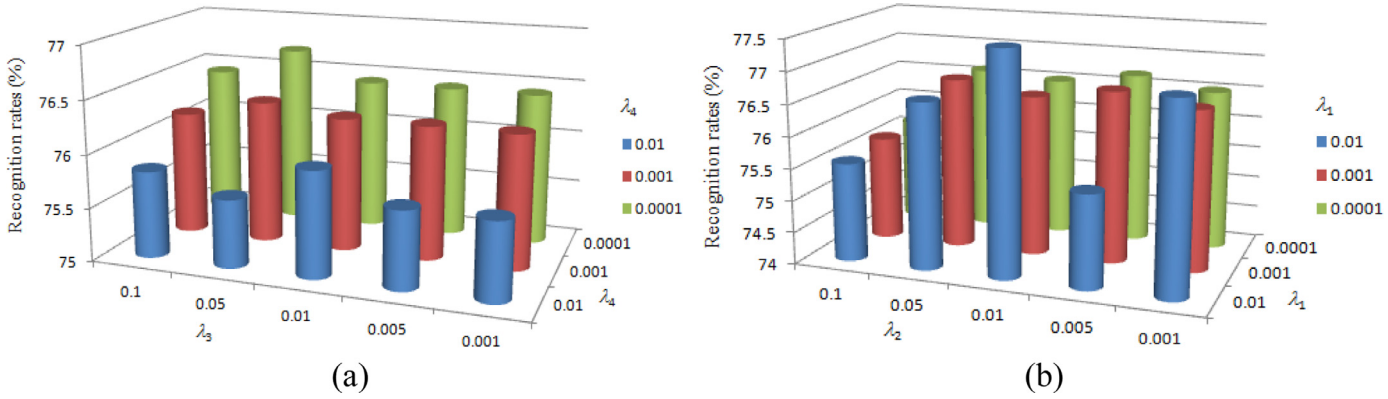

Fig. 8. The recognition rates versus parameters on the LFW database. (a) The recognition rates versus λ_3 and λ_4. (b) The recognition rates versus λ_1 and λ_2.

Fig. 9. The first line: images of one person from subset 1. The second line: images of one person from subset 3. The third line: images of one person from subset 4.

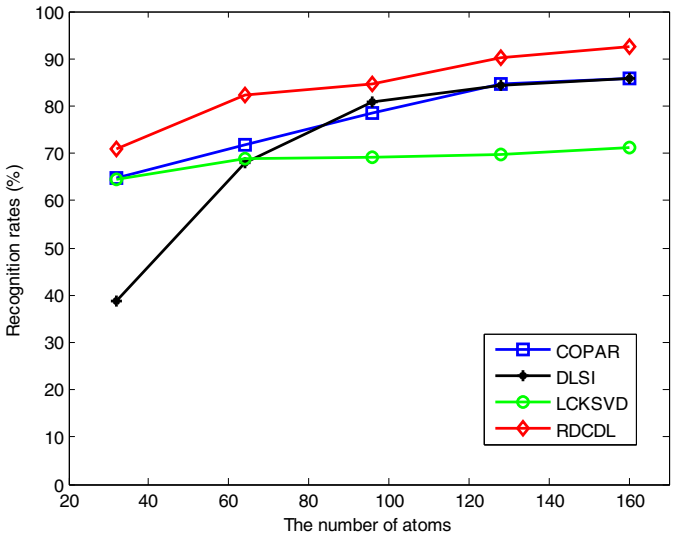

Fig. 10. The recognition rates versus the number of atoms on the Extended Yale B database.

higher than that of ESRC when the feature dimension of images is 100, 130 and 150, respectively. RDCDL takes less training time than all the compared algorithms except LDL. The training speed of RDCDL is 1.4 times and 3.5 times faster than that of DLSI and FDDL, respectively. RDCDL takes a small testing time similar to all the compared algorithms. In order to further evaluate the performance of the proposed algorithm, we compare it with some dictionary learning algorithms such as COPAR, DLSI and LCKSVD. Fig. 10 shows the recognition rates of COPAR, DLSI, LCKSVD and RDCDL with the different number of atoms ($K = 32, 64, 96, 128, 160$) when the feature dimension of images is 150. From Fig. 10, we can see that the recognition rate of RDCDL is much higher than that of the compared algorithms. For example, when the number of atoms is 160, the recognition rate of RDCDL is at least 6% higher than that of the compared algorithms.

6.3. Experimental results on the AR database

The AR database consists of over 4,000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separated sessions. Fig. 11 shows images of one person from the AR face database. As in [5], we choose a subset consisting of 50 male subjects and 50 female subjects in the experiment. The size of the original face images is 165×120. In the experiment, we randomly select 90 subjects from session 1 for training and the same 90 subjects from session 2 for testing. For each subject, the 7 images with illumination and expression change from session 1 are used for training, the 13 images with illumination, expression change, sunglasses and scarf from session 2 are used for testing. Because there are drastic illumination and expression changes in the face images of the AR database, the alternative training samples are produced by mirroring the original training samples. Here the parameters of sparse constraint and discriminative representation are set as $\lambda_2 = 0.005$ and $\lambda_3 = 0.05$. In order to construct the set of disturbance components, the remainder 10 subjects from session 1 are selected. There are 13 images with illumination, expression change, sunglasses and scarf per subject in session 1. The disturbance components are computed by Eq. (7). The number of dictionary atoms is the same as the number of original training samples and the feature dimension of images is reduced to 400, 500 and 600 via PCA.

Table 4 shows the recognition results of the proposed algorithm and ten compared algorithms. When the feature dimension of images is 600, the computing time of training dictionaries and classifying a testing sample is also shown in Table 4. From Table 4, we can see that RDCDL achieves the highest recognition rates among all the algorithms. The recognition of RDCDL is at least 15% higher than that of the compared algorithms except ESRC. ESRC is the second best algorithm. Moreover, the recognition rate of RDCDL is 1.4%, 2.0% and 1.8% higher than that of ESRC when the feature di-

Fig. 11. The first line: images of one person from session 1. The second line: images of one person from session 2.

Table 4
The recognition rates (%) and computing time for training dictionaries and classifying a testing sample on the AR database.

Algorithm	400	500	600	TrT (s)	TtT (s)
SVM	38.7	38.8	38.9	–	1.2e–3
SRC[5]	70.1	70.2	70.7	–	4.2e–2
CRC[43]	70.1	70.3	70.7	–	1.6e–3
DKSVD[10]	66.7	68.1	68.8	1339.4	2.3e–1
LCKSVD[11]	67.6	69.1	69.7	21.9	2.5e–3
COPAR[12]	66.8	67.0	67.5	464.5	1.4e–3
FDDL[21]	71.1	70.7	71.0	993.9	4.6e–2
DLSI[20]	71.2	71.5	71.5	258.4	4.3e–2
LDL[42]	69.7	70.1	70.4	16.1	5.9e–2
ESRC[25]	85.2	85.4	85.6	–	4.6e–2
RDCDL	**86.6**	**87.4**	**87.4**	105.3	6.4e–2

Table 5
The recognition rates (%) and computing time for training dictionaries and classifying a testing sample on the Multi-PIE database.

Algorithm	210	240	270	TrT (s)	TtT (s)
SVM	55.2	55.2	55.2	–	3.3e–4
SRC[5]	91.5	91.3	91.7	–	9.7e–3
CRC[43]	91.5	90.5	91.7	–	6.5e–4
DKSVD[10]	88.8	87.8	89.2	152.6	4.4e–2
LCKSVD[11]	88.8	88.2	88.2	22.8	9.3e–4
COPAR[12]	88.0	87.8	87.5	26.4	1.2e–3
FDDL[21]	91.8	91.7	91.8	156.6	1.2e–2
DLSI[20]	90.8	90.8	91.2	28.2	9.9e–3
LDL[42]	90.8	91.0	91.3	7.2	1.1e–2
ESRC[25]	92.3	92.2	92.5	–	1.1e–2
RDCDL	**92.5**	**94.2**	**95.7**	8.6	1.2e–2

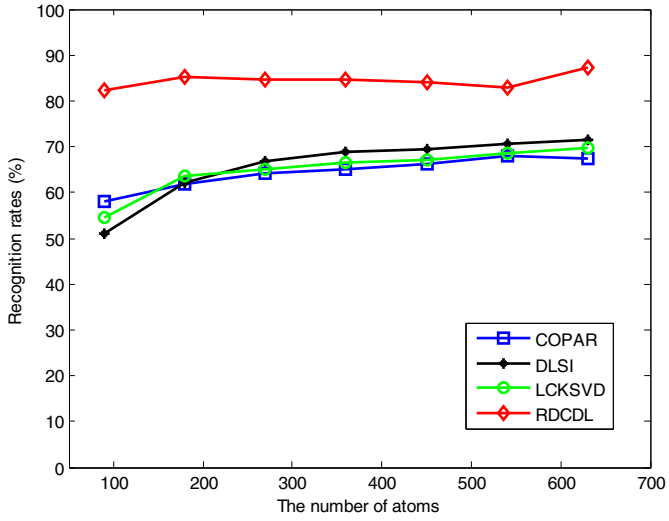

Fig. 12. The recognition rates versus the number of atoms on the AR database.

mension of images is 400, 500 and 600, respectively. RDCDL takes less training time than DKSVD, COPAR, FDDL and DLSI. For example, the training speed of RDCDL is 12.7 times and 9.4 times faster than that of DKSVD and FDDL, respectively. RDCDL takes more training time than LCKSVD and LDL. RDCDL takes a small testing time similar to all the compared algorithms except DKSVD. In order to test the performance of the proposed algorithm, we also compare it with some dictionary learning algorithms such as COPAR, DLSI and LCKSVD. Fig. 12 shows the recognition rates of COPAR, DLSI, LCKSVD and RDCDL with the different number of atoms ($K = 90, 180, 270, 360, 450, 540, 630$) when the feature dimension of images is 600. From Fig. 12, we can see clearly that RDCDL significantly outperforms the compared algorithms, and the recognition rate of RDCDL is more than 12% higher than that of the compared algorithms when the number of atoms is from 90 to 630.

6.4. Experimental results on the Multi-PIE database

The CMU Multi-PIE face database is a large scale database of 337 subjects including four sessions with simultaneous variations of pose, expression and illumination. Fig. 13 shows some images of one person from the Multi-PIE face database. Among the 337 subjects, we choose the former 60 subjects from session 1 as the training set and the same subjects from session 3 as the testing set. For each subject, we choose the 5 frontal images with illumination {0, 1, 3, 4, 6} and smile expression from session 1 for training and the 10 frontal images with illumination {0, 2, 4, 6, 8, 10, 12, 14, 16, 18} and smile expression from session 3 for testing. Due to the mild variations of illumination in the face images of the Multi-PIE database, the alternative training samples are produced by corrupting the original training samples using the salt & pepper noise, whose density is 0.5. In order to construct the set of disturbance components, we select 10 subjects different to the 60 training subjects from session 1, each of these 10 subjects has the 5 frontal images with illumination {1, 3, 7, 11, 13} and smile expression. The disturbance components are computed by Eq. (7). The number of dictionary atoms is the same as the number of original training samples and the feature dimension of images is reduced to 210, 240 and 270 via PCA.

Table 5 shows the recognition results of the proposed algorithm and ten compared algorithms. When the feature dimension of images is 270, the computing time of training dictionaries and classifying a testing sample is also shown in Table 5. As shown in Table 5, we can see that the recognition rate of RDCDL is higher than that of the compared algorithms. ESRC achieves the second best recognition rates. FDDL achieves the third best recognition rates. The recognition rate of RDCDL is 0.7%, 2.5% and 3.9% higher than that of FDDL when the feature dimension of images is 210, 240 and 270, respectively. What's more, the recognition rate of RDCDL is 2.0% and 3.2% higher than that of ESRC when the feature dimension of images is 240 and 270, respectively. RDCDL takes less training time than all the compared algorithms except LDL. The training speed of RDCDL is 2.7 times faster than that of LCKSVD,

Fig. 13. Top: some images of one person from session 1. Bottom: some images of one person from session 2.

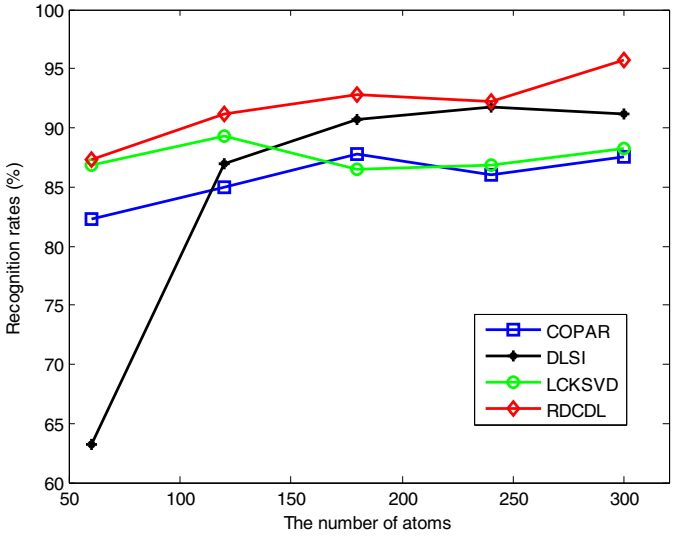

Fig. 14. The recognition rates versus the number of atoms on the Multi-PIE database.

which is the third fastest algorithm. RDCDL takes more testing time than all the compared algorithms except DKSVD and FDDL. The testing time of RDCDL is equal to that of FDDL. The testing speed of RDCDL is 3.7 times faster than that of DKSVD. Fig. 14 shows the recognition rates of COPAR, DLSI, LCKSVD and RDCDL with the different number of atoms ($K = 60, 120, 180, 240, 300$) when the feature dimension of images is 270. From Fig. 14, it can be seen that RDCDL outperforms the compared algorithms.

6.5. Experimental results on the FRGC database

In this experiment, we perform the test on the FRGC 4-th experiment, which has a target set with 16,028 samples and a query set with 8014 samples collected from 466 subjects. The samples in the target set were captured under controlled illumination, while the samples in the query set were captured uncontrolled illumination. Fig. 15 shows some images of one person from the FRGC face database. We select the former 100 subjects and the former 10 images per subject from the target set for training and the same 100 subjects from the query set for testing. There are relatively drastic illumination changes in the face images of the FRGC database, so the alternative training samples are produced by occluding the original training samples using the random square block occlusion, whose level is 0.05. In order to construct the set of disturbance components, we directly use 10 subjects for constructing the set of disturbance components of the experiment on the Multi-PIE database. Each of these 10 subjects has the 10 frontal images with illumination {0, 1, 3, 4, 6, 7, 8, 11, 13, 14} and smile expression. The disturbance components are computed by Eq. (7). When the feature dimension of images is reduced to 300, 400 and 500 via PCA, the recognition results of the proposed algorithm and the compared algorithms are listed in Table 6.

Table 6
The recognition rates (%) and computing time for training dictionaries and classifying a testing sample on the FRGC database.

Algorithm	300	400	500	TrT (s)	TtT (s)
SVM	7.2	7.2	7.2	–	1.5e−3
SRC[5]	25.7	25.8	26.5	–	6.0e−2
CRC[43]	25.9	25.9	25.4	–	2.4e−3
DKSVD[10]	16.2	19.6	20.3	2934.8	3.5e−1
LCKSVD[11]	19.0	20.0	21.0	71.9	2.3e−3
COPAR[12]	19.8	19.9	20.0	909.8	1.2e−3
FDDL[21]	25.9	26.5	26.7	3000.3	5.9e−2
DLSI[20]	21.4	22.4	23.8	249.1	6.9e−2
LDL[42]	23.0	23.4	23.8	30.4	5.2e−2
ESRC[25]	25.2	26.0	26.2	–	6.3e−2
RDCDL	**32.9**	**34.2**	**35.9**	139.8	1.6e−1

When the feature dimension of images is 500, Table 6 also shows the computing time of training dictionaries and classifying a testing sample. From Table 6, we can see that the recognition rate of RDCDL improves at least 7% over the compared algorithms. RDCDL takes less training time than DKSVD, COPAR, FDDL and DLSI. The training speed of RDCDL is 21.0, 6.5, 21.5 and 1.8 times faster than that of DKSVD, COPAR, FDDL and DLSI, respectively. RDCDL takes more testing time than all the compared algorithms except DKSVD. RDCDL takes less testing time than DKSVD. The testing speed of RDCDL is 2.2 times faster than that of DKSVD. When the feature dimension of images is 500, Fig. 16 shows the recognition rates of COPAR, DLSI, LCKSVD and RDCDL with the different number of atoms ($K = 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000$). From Fig. 16, it can be seen that the recognition rate of RDCDL is much higher than that of the compared algorithms. When the number of atoms is 1000, the recognition rate of RDCDL is about 16%, 15% and 12% higher than that of COPAR, LCKSVD and DLSI, respectively.

6.6. Experimental results on the LFW database

RDCDL is evaluated on the application of face recognition in the wild. The aligned labeled face in the wild (LFWa) is used here. LFW is a large-scale database, which contains variations of pose, illumination, expression, misalignment and occlusion, etc. Fig. 17 shows images of one person from LFWa. In the experiment, 136 subjects with no less than 11 samples per subject are chosen, we choose the former 100 subjects from 136 subjects. For each subject, we choose the former 10 images for training and the remainder images for testing. There are variations of pose and misalignment in the face images of the LFW database, so the alternative training samples are produced by mirroring the original training samples. In order to construct the set of disturbance components, the remainder 36 subjects from LFWa are selected. The disturbance components are computed by Eq. (7). Histogram of Uniform-LBP is extracted via dividing a face image into 10×8 patches. When the di-

Fig. 15. Top: some images of one person from the target set. Bottom: some images of one person from the query set.

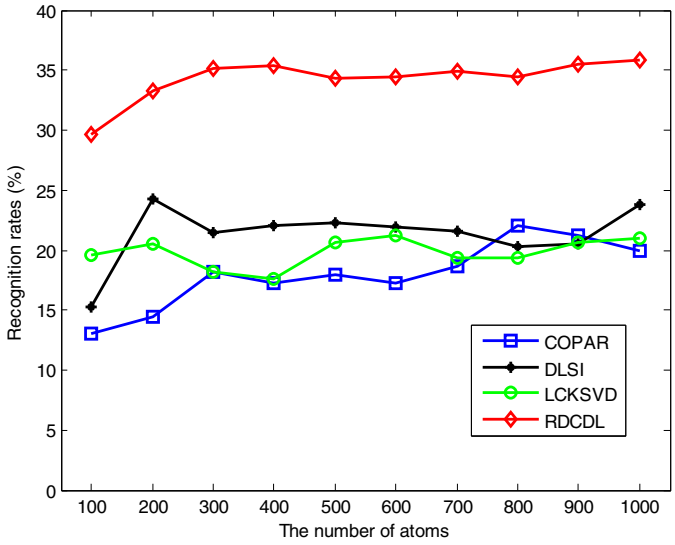

Fig. 16. The recognition rates versus the number of atoms on the FRGC database.

mension of histograms is reduced to 300, 400 and 500 via PCA, Table 7 shows the recognition results of the proposed algorithm and ten compared algorithms.

When the dimension of histograms is 500, Table 7 also shows the computing time of training dictionaries and classifying a testing sample. From Table 7, we can see that RDCDL achieves the highest recognition rates among all the algorithms. FDDL achieves the second highest recognition rates. When the dimension of histograms is 300, 400 and 500, the recognition rate of RDCDL is 2.6%, 3.7% and 2.6% higher than that of FDDL, respectively. RDCDL takes less training time than DKSVD, COPAR, FDDL and DLSI. The training speed of RDCDL is 3.7, 2.9, 2.9 and 1.8 times faster than that of DKSVD, COPAR, FDDL and DLSI, respectively. RDCDL takes more testing time than all the compared algorithms. Fig. 18 shows the recognition rates of COPAR, DLSI, LCKSVD and RDCDL with the different number of atoms ($K = 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000$) when the dimension of histograms is 500. As shown in Fig. 18, we can see that the RDCDL outperforms the compared algorithms.

Table 7
The recognition rates (%) and computing time for training dictionaries and classifying a testing sample on the LFW database.

Algorithm	300	400	500	TrT (s)	TtT (s)
SVM	51.5	53.1	53.9	–	2.2e−3
SRC[5]	67.9	70.0	70.8	–	4.0e−2
CRC[43]	69.7	71.1	72.2	–	2.2e−3
DKSVD[10]	61.7	64.8	67.9	654.3	2.7e−2
LCKSVD[11]	62.4	65.1	67.3	157.1	7.2e−4
COPAR[12]	63.1	65.6	68.6	517.9	2.2e−3
FDDL[21]	71.8	73.3	74.1	501.8	5.2e−2
DLSI[20]	70.7	72.8	73.6	309.3	4.9e−2
LDL[42]	71.8	72.1	73.4	10.9	7.1e−2
ESRC[25]	70.3	71.0	73.1	–	7.6e−2
RDCDL	**74.4**	**77.0**	**76.7**	175.6	2.2e−1

6.7. Experimental analysis

The above experiments show that the proposed RDCDL achieves higher recognition rates than SVM, SRC and CRC, which directly use the original training samples for face recognition. This clearly demonstrates that the learned dictionaries have more discriminative ability than the original training samples. The experiments show that the recognition rate of the proposed RDCDL is higher than that of DKSVD, LCKSVD, COPAR, FDDL, DLSI and LDL. Although they all aim to learn the dictionary from the training data, the experiments effectively demonstrate that the proposed RDCDL has more power discriminative ability than them. The experiments also show that the proposed RDCDL outperforms ESRC, which directly uses the original training samples and the disturbance components for face recognition. This demonstrates that the learned dictionar-

Fig. 17. Images of one person from LFWa.

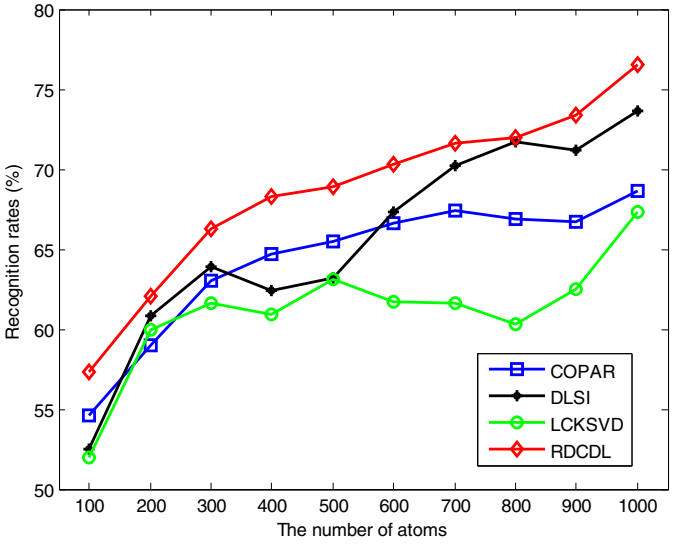

Fig. 18. The recognition rates versus the number of atoms on the LFW database.

ies have more discriminative ability than the original training samples and the disturbance components.

The experiments show that the proposed RDCDL takes less training time than DKSVD, COPAR, FDDL and DLSI and slightly more training time than LDL. Although the experiments show that the proposed RDCDL takes more testing time than all the compared algorithms except FDDL and DKSVD, the testing time of the proposed RDCDL is less than 0.3 s, RDCDL can be applied to the practical face recognition.

6.8. Discussion on RDCDL with deep features

Recently deep learning has been used to learn deep features for the face image and achieved good performances in face recognition [51-55]. Sun et al. [51] proposed to learn the high-level deep features called Deep hidden IDentity features (DeepID) by using the deep convolutional neural networks for face verification. Sun et al. [52] proposed to learn the Deep IDentification-verification features (DeepID2) by using the well-designed deep convolutional networks, with the human-level accuracy achieved in face recognition. Taigman et al. [53] used the identity labeled dataset of four million facial images belonging to more than 4000 identities to train a deep network to achieve the approximate human-level performance in face verification. Lu et al. [54] proposed a discriminative deep metric learning method to train a deep neural network for face and kinship verification in wild conditions. Li et al. [55] proposed a distance metric optimization driven deep learning model for age invariant face recognition, which learned features and a distance metric simultaneously.

Obviously the proposed RDCDL using deeps feature will possibly achieve better performance in face recognition. For instance, deep convolutional neural network features can be extracted from the original training samples and the samples with simulated variations. The powerful deep features have more discriminative ability for face representation, which will have better performance than the hand-crafted features. For the extraction of real variation, one possible way is to extract the deep features of the generic training samples first, and then to extract the deep feature matrix of real variations via Eq. (7). Since we mainly focus on the design of robust classifiers, in this paper we proposed the model of RDCDL based on the hand-craft features for face recognition. We will develop the model of RDCDL with deep features in the future.

7. Conclusion

In the paper, we propose a new robust, discriminative and comprehensive dictionary learning (RDCDL) model. The proposed model uses the extracted real facial variation and the alternative training samples which are produced by various schemes to obtain the robust dictionary. The proposed model has learned the dictionary including class-shared dictionary atoms, class-specific dictionary atoms and disturbance dictionary atoms to completely represent the commonality, particularity and disturbance components in the data belonging to different classes. The discrimination of the dictionary and the representation coefficients is exploited via the designed discriminative regularizations and it effectively improves the classification capability of the learned dictionary. Extensive experiments on face recognition have demonstrated the effectiveness of RDCDL to those state-of-the-art methods. The proposed RDCDL not only can be used for face recognition but also can be applied to other pattern classification. In the future, we will apply RDCDL to other pattern classification.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Grant nos. 61772568, U1713208, 61472187, and 61672357), Guangzhou Science and Technology Program (Grant no. 201804010288), the 973 Program No. 2014CB349303, Program for Changjiang Scholars, Shenzhen Scientific Research and Development Funding Program (Grant no. JCYJ20170302153827712), the Scientific Research Project of Sichuan University of Science and Engineering (Grant nos. 2015RC16 and 2015RC49), the Open Fund Project of Artificial Intelligence Key Laboratory of Sichuan Province (Grant nos. 2015RZY01 and 2016RYY02), the Project of Sichuan Provincial Department of Education (Grant nos. 17ZB0302 and 14ZA0202).

References

[1] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (12) (2006) 3736–3745.
[2] J.C. Yang, J. Wright, Y. Ma, T. Huang, Image super-resolution as sparse representation of raw image patches, in: Proceedings of the CVPR, 2008.
[3] J.C. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the CVPR, 2009.
[4] Y. Xu, B. Zhang, Z. Zhong, Multiple representations and sparse representation for image classification, Pattern Recognit. Lett. 68 (2015) 9–14.
[5] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[6] Y. Xu, Z. Zhang, G.M. Lu, J. Yang, Approximately symmetrical face images for image preprocessing in face recognition and sparse representation based classification, Pattern Recognit. 54 (2016) 68–82.
[7] Y. Liu, X.M. Li, C.Y. Liu, H.X. Liu, Structure-constrained low-rank and partial sparse representation with sample selection for image classification, Pattern Recognit. 59 (2016) 5–13.
[8] J. Mairal, F. Bach, J. Ponce, Task-driven dictionary learning, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 791–804.
[9] M. Yang, L. Zhang, S.C.K Shiu, D. Zhang, Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary, Pattern Recognit. 46 (2013) 1865–1878.
[10] Q. Zhang, B.X. Li, Discriminative K-SVD for dictionary learning in face recognition, in: Proceedings of the CVPR, 2010.
[11] Z.L. Jiang, Z. Lin, L.S. Davis, Label consistent K-SVD: Learning a discriminative dictionary for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2651–2664.
[12] S. Kong, D.H. Wang, A dictionary learning approach for classification: separating the particularity and the commonality, in: Proceedings of the ECCV, 2012.
[13] N. Zhou, J.P. Fan, Learning inter-related visual dictionary for object recognition, in: Proceedings of the CVPR, 2012.
[14] Z.Z. Feng, M. Yang, L. Zhang, Y. Liu, D. Zhang, Joint discriminative dimensionality reduction and dictionary learning for face recognition, Pattern Recognit. 46 (2013) 2134–2143.
[15] X.Y. Jing, F. Wu, X.K. Zhu, X.W. Dong, F. Ma, Z.Q. Li, Multi-spectral low-rank structured dictionary learning for face recognition, Pattern Recognit. 59 (2016) 14–25.

[16] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing over complete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[17] M. Yang, L. Zhang, J. Yang, D. Zhang, Metaface learning for sparse representation based face recognition, in: Proceedings of the ICIP, 2010.
[18] P. Sprechmann, G. Sapiro, Dictionary learning and sparse coding for unsupervised clustering, in: Proceedings of the International Conference on Acoustics Speech and Signal Processing, 2010.
[19] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Supervised dictionary learning, in: Proceedings of the NIPS, 2009.
[20] I. Ramirez, P. Sprechmann, G. Sapiro, Classification and clustering via dictionary learning with structured incoherence and shared features, in: Proceedings of the CVPR, 2010.
[21] M. Yang, L. Zhang, X.C. Feng, D. Zhang, Fisher discrimination dictionary learning for sparse representation, in: Proceedings of the ICCV, 2011.
[22] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zissserman, Learning discriminative dictionaries for local image analysis, in: Proceedings of the CVPR, 2008.
[23] K.K. Huang, D.Q. Dai, C.X. Ren, Z.R. Lai, Learning kernel extended dictionary for face recognition, IEEE Trans. Neural Learn. Syst. 1 (2016) 1–13.
[24] W.H. Ou, X.G. You, D.C. Tao, P.Y. Zhang, Y.Y. Tang, Z.Q. Zhu, Robust face recognition via occlusion dictionary learning, Pattern Recognit. 47 (2014) 1559–1572.
[25] W.H. Deng, J.N. Hu, J. Guo, Extended SRC: Undersampled face recognition via intraclass variation dictionary, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1864–1870.
[26] A. Castrodad, G. Sapiro, Sparse modeling of human actions from motion imagery, Int. J. Comput. Vis. 100 (2012) 1–15.
[27] L. Shen, S.H. Wang, G. Sun, S.Q. Jiang, Q.M. Huang, Multi-level discriminative dictionary learning towards hierarchical visual categorization, in: Proceedings of the CVPR, 2013.
[28] R. Rubinstein, A.M. Bruckstein, M. Elad, Dictionaries for sparse representation modeling, Proc. IEEE 98 (6) (2010) 1045–1057.
[29] Y.G. Peng, A. Ganesh, J. Wright, Y. Ma, RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images, in: Proceedings of the CVPR, 2010.
[30] L. Rosasco, A. Verri, M. Santoro, S. Mosci, S. Villa, Iterative projection methods for structured sparsity regularization, MIT Technical reports, MIT-CSAIL-TR-2009-050 (2009) CBCL-282.
[31] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.
[32] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[33] A. Martinez, R. Benavente, The AR face database, CVC Technical report No. 24, 1998.
[34] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (2010) 807–813.
[35] L. Wolf, T. Hassner, Y. Taigman, Similarity scores based on background samples, in: Proceedings of the (ACCV), 2009.
[36] M. Yang, L. Zhang, X.C. Feng, D. Zhang, Sparse representation based Fisher discrimination dictionary learning for image classification, Int. J. Comput. Vis. 109 (2014) 209–232.
[37] H.Y. Chang, M. Yang, J. Yang, Learning a structure adaptive dictionary for sparse representation based classification, Neurocomputing 190 (2016) 124–131.
[38] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: Proceedings of the (CVPR), 2005.
[39] C.P. Po, Y.C.F Wang, Undersampled face recognition via robust auxiliary dictionary learning, IEEE Trans. Image Process. 24 (6) (2015) 1722–1734.
[40] S. Shekhar, V.M. Patel, R. Chellappa, Analysis sparse coding models for image-based classification, in: Proceedings of the (ICIP), 2014.
[41] S. Gu, L. Zhang, W. Zuo, X. Feng, Projective dictionary pair learning for pattern classification, in: Proceedings of the (NIPS), 2014.
[42] M. Yang, D.X. Dai, L.L. Shen, L.V. Gool, Latent dictionary for sparse representation based classification, in: Proceedings of the (CVPR), 2014.
[43] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: which helps face recognition, in: Proceedings of the (ICCV), 2011.
[44] S.J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, A interior-point method for large-scale l_1-regularized least squares, IEEE J. Sel. Top. Signal Process. 1 (2007) 606–617.
[45] J.R. Shewchuk, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, School of Computer Science Carnegie Mellon University Pittsburgh, PA, Technical Reports, 1994 CMU-CS-94-125.
[46] J. Guo, Y.Q. Guo, X.W. Kong, M. Zhang, R. He, Discriminative analysis dictionary learning, in: Proceedings of the (AAAI), 2016.
[47] R. Rubinstein, M. Elad, Dictionary learning for analysis synthesis thresholding, IEEE Trans. Signal Process. 62 (22) (2014) 5962–5972.
[48] M. Yang, W.Y. Liu, W.X. Luo, L.L. Shen, Analysis-synthesis dictionary learning for universality-particularity representation based classification, Proceedings of the (AAAI), 2016.
[49] F. Wu, X.Y. Jing, X.G. You, D. Yue, R.M. Hu, J.Y. Yang, Multi-view low-rank dictionary learning for image classification, Pattern Recognit. 50 (2016) 143–154.
[50] Y. Xu, Z.M. Li, B. Zhang, J. Yang, J. You, Sample diversity, representation effectiveness and robust dictionary learning for face recognition, Inf. Sci. 375 (2017) 171–182.
[51] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10000 classes, in: Proceedings of the (CVPR), 2014.
[52] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of the (NIPS), 2014.
[53] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: Proceedings of the (CVPR), 2014.
[54] J.W. Lu, J.L. Hu, Y.P. Tan, Discriminative deep metric learning for face and kinship verification, IEEE Trans. Image Process. 26 (9) (2017) 4269–4282.
[55] Y. Li, G.R. Wang, L. Nie, Q. Wang, W.W. Tan, Distance metric optimization driven convolutional neural network for age invariant face recognition, Pattern Recognit. 75 (2018) 51–62.
[56] M. Yang, H. Chang, W. Luo, Discriminative analysis-synthesis dictionary learning for image classification, Neurocomputing 219 (2017) 404–411.
[57] M. Yang, H. Chang, W. Luo, J. Yang, Fisher discrimination dictionary pair learning for image classification, Neurocomputing 269 (2017) 13–20.

Guojun Lin received the BS degree from Zhejiang University of Technology in 2001, the MS degree from Southwest Jiaotong University in 2008, and the Ph.D. degree from University of Electronic Science and Technology of China in 2014. From May to October in 2008, he was a software engineer at Skypine Electronics (Shenzhen) Co., Ltd. He is currently a lecturer at School of Automation and Electric Information, Sichuan University of Science and Engineer and a research assistant at School of Computer Science and Software Engineering, Shenzhen University. His research interest includes sparse coding, dictionary learning and face recognition.

Meng Yang is currently an associate professor at School of Data and Computer Science, Sun Yat–Sen University, Guangzhou, China. He received his Ph.D. degree from The Hong Kong Polytechnic University in 2012. Before joining Shenzhen University, he has been working as Postdoctoral fellow in the Computer Vision Lab of ETH Zurich. His research interest includes sparse coding, dictionary learning, object recognition and machine learning. He has published 10 AAAI/CVPR/ICCV/ICML/ECCV papers and several IJCV, IEEE TNNLS and TIP journal papers.

Jian Yang received the B.S. degree in mathematics from the Xuzhou Normal University in 1995. He received the M.S. degree in applied mathematics from the Changsha Railway University in 1998 and the Ph.D. degree from the Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a professor in the School of Computer Science and Technology of NUST. He is the author of more than 80 scientific papers in pattern recognition and computer vision. His journal papers have been cited more than 3000 times in the ISI Web of Science, and 7000 times in the Web of Scholar Google. His research interests include pattern recognition, computer vision and machine learning. Currently, he is an associate editor of Pattern Recognition Letters and IEEE Trans. Neural Networks and Learning Systems, respectively.

Linlin Shen received the B.Sc. degree from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree from University of Nottingham, Nottingham, U.K., in 2005. He was a Research Fellow with the Medical School, University of Nottingham, researching brain image processing of magnetic resonance imaging. He is currently a Professor and the Director of the Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, China. His current research interests include Gabor wavelets, face/palmprint recognition, medical image processing, and image classification.

Weicheng Xie received the B.S. degree in statistics from Central China Normal University in 2008, the M.S. degree in probability and mathematical statistics and Ph.D. degree in computational mathematics from Wuhan University, China in 2010 and 2013. He has been a visiting research fellow with School of Computer Science, University of Nottingham, UK. He is now a postdoctoral researcher at Shenzhen Key Laboratory of Spatial Smart Sensing and Services, School of Computer Science and Software Engineering, Shenzhen University. His current researches focus on image processing and facial expression analysis.