Multimed Tools Appl (2017) 76:8969–8990
DOI 10.1007/s11042-016-3492-1

Simultaneous dimensionality reduction and dictionary learning for sparse representation based classification

Bao-Qing Yang [1,2] · Chao-Chen Gu [1,2] · Kai-Jie Wu [1,2] · Tao Zhang [1,2] · Xin-Ping Guan [1,2]

Received: 14 September 2015 / Revised: 20 January 2016 / Accepted: 22 March 2016 /
Published online: 15 April 2016

Abstract Learning dictionaries from the training data has led to promising results for pattern classification tasks. Dimensionality reduction is also an important issue for pattern classification. However, most existing methods perform dimensionality reduction (DR) and dictionary learning (DL) independently, which may result in not fully exploiting the discriminative information of the training data. In this paper, we propose a simultaneous dimensionality reduction and dictionary learning (SDRDL) model to learn a DR projection matrix and a class-specific dictionary (i.e., the dictionary atoms correspond to the class labels) simultaneously. Since simultaneously learning makes the learned projection and dictionary fit better with each other, more effective pattern classification can be achieved using the representation residual. In SDRDL model, not only the representation residual is discriminative, but the representation coefficients are also discriminative. Therefore, a classification scheme associated with SDRDL is presented by exploiting such discriminative information. Experimental results on a series of benchmark image databases show that our proposed method outperforms many state-of-the-art discriminative dictionary learning methods.

Keywords Dictionary learning · Sparse representation · Dimensionality reduction · Image classification

✉ Bao-Qing Yang
yang_baoqing@sjtu.edu.cn

✉ Chao-Chen Gu
jacygu@sjtu.edu.cn

[1] Department of Automation, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, People's Republic of China

[2] Key Laboratory of System Control and Information Processing, Ministry of Education of China, 800 Dongchuan Road, Shanghai, People's Republic of China

1 Introduction

With the inspiration of sparse mechanism of human vision system [28, 29], sparse representation has become an appealing concept for data representation and achieved competitive performance in image restoration [6, 10, 11, 21] and compressed sensing [8]. Sparse representation can also be used for effective image classification, e.g. face recognition (FR) [40, 42, 45, 48, 52], digit and texture classification [16, 24, 33, 46], etc.

Sparse representation based classification (SRC) framework involves two steps: coding and classification. First, over a dictionary of atoms, a query signal/image is collaboratively coded with some sparse constraint. Then, classification is performed by using the coding coefficients and the dictionary. The dictionary for sparse coding could be predefined. For example, Wright et al. [42] populated a dictionary with the original training samples from all classes to code the query face image, and classified the query face image by passing the coding coefficients into a minimum reconstruction error classifier. This so called SRC classifier shows very competitive performance, but the dictionary adopted in it is not effective enough to represent the query images for performing classification due to two issues. One is that taking the original training samples as the dictionary could not effectively exploit the discriminative information in the training samples. The other is that taking analytically designed off-the-shelf bases as dictionary (e.g., [16] takes Haar wavelets and Gabor wavelets as the dictionary) might be effective enough for universal types of images but not for specific type of images such as face, digit and texture images. These two issues of predefined dictionary can be addressed, to a certain extent, by properly learning a desired dictionary from the given original training samples.

Dictionary learning (DL) devotes to learning from training samples the optimal dictionary over which the given signal could be well represented or coded for processing. A number of DL methods have been proposed for image restoration [1, 6, 10, 11, 25, 55] and image classification [9, 14, 18, 19, 22, 24, 25, 31–34, 41, 46, 48–50, 52]. One representative DL method for image restoration is K-means singular value decomposition (KSVD) algorithm [1], which learns an over-complete dictionary from example natural images and uses this learned dictionary for image restoration. Inspired by KSVD, many reconstructive dictionary learning (DL) methods [6, 10, 11, 25, 55] have been proposed and showed state-of-the-art performance in image restoration tasks. Although these reconstructive dictionary learning (DL) methods have achieved promising results in image restoration, they are not favorable for image classification since their objective is only to represent the training samples faithfully. Different from image restoration, image classification aims to classify the input query sample correctly. Therefore, the discriminative ability of the learned dictionary is the major concern for image classification. Up to now, Discriminative DL methods have been proposed to promote the discriminative ability of the learned dictionary [9, 14, 18, 19, 22, 24, 25, 31–34, 41, 46, 48–50, 52] and have led to many state-of-the-art results in pattern classification problems.

One popular type of discriminative DL methods aims to learn a shared dictionary for all classes while improving the discriminative ability of the coefficients vector. In the DL methods proposed by Rodriguez et al. [34] and Jiang et al. [19], the coefficients vector of the samples from the same class are encouraged to be similar to each other. As well as the l_0- or l_1-norm sparsity penalty, nonnegative [15], group [4, 38] and structured sparsity penalty [17] have been proposed to be imposed on the representation coefficients in different applications. It is popular concurrently to learn a dictionary and a classifier over the coefficients vector. In this spirit, Mairal et al. [24] and Pham et al. [25] proposed to learn discriminative dictionaries while

training linear classifiers jointly. Inspired by the work of Pham et al. [25], Zhang et al. [52] extended the original K-SVD algorithm by simultaneously learning a linear classifier for face recognition. Following Zhang et al. [52], Jiang et al. [19] proposed Label-Consistent KSVD by introducing a label consistent regularization to enforce the discrimination of coding vectors. The so-called LC-KSVD algorithm exhibits good classification results. Recently, Mairal et al. [25] proposed a task-driven DL (TDDL) framework in which different risk functions of the representation coefficients are minimized for different tasks. Cai et al. [7] proposed a parameterization method to adaptively determine the weight of each coding vector pair, which leads to a support vector guided dictionary learning (SVGDL) model.

Another type of discriminative DL methods aims to learn a class-specific dictionary whose atoms correspond to the subject class labels. Mairal et al. [22] modified the KSVD model by introducing a discriminative reconstruction penalty term for texture segmentation and scene analysis. Yang et al. [47] and Sprechmann et al. [37] learned a dictionary for each class with sparse coefficients and used the learned dictionary for face recognition and signal clustering, respectively. Castrodad et al. [9] proposed to impose non-negative penalty on both dictionary atoms and representation coefficients to learn a set of action-specific dictionaries. From the training images of each category, Wu et al. [44] learned active basis models for object detection and recognition. Ramirez et al. [33] encouraged the dictionaries associated with different classes to be as independent as possible by introducing an incoherence promoting term. Following Ramirez et al. [33], Wang et al. [41] presented a class-specific DL algorithm for sparse modeling in action recognition. Yang et al. [49, 50] proposed to learn a structural dictionary by imposing the Fisher discrimination criterion on the sparse coding coefficients to enhance class discrimination power.

On the other, numerous dimensionality reduction methods are developed for feature extraction. The representative dimensionality reduction methods include Principal Component Analysis (PCA) [39], Linear Discriminate Analysis (LDA) [3] and Locality Preserving Projection (LPP) [27], etc. In all the previous DL methods, the dimensionality reduction (DR) and dictionary learning (DL) are studied as two independent processes. Traditionally, DR is performed first to the training samples and the reduced dimensionality feature are used for DL. However, the pre-learned DR projection may not preserve the best feature for DL. Therefore, the DR and DL processes should be simultaneously conducted for a more effective classification task. Some works has been done to investigate the dimensionality reduction for DL. For example, Zhang et al. [53] proposed an unsupervised learning method for dimensionality reduction in SRC. Feng et al. [12] jointly trained a dimensionality reduction transform and a dictionary for face recognition. Both of these two methods show higher FR rates than PCA and random projection.

In this paper, we propose a simultaneous dimensionality reduction and dictionary learning (SDRDL) model to learn a dimensionality reduction (DR) projection matrix P and a class-specific dictionary D (i.e., the dictionary atoms have correspondences to the class labels) simultaneously for pattern classification. In the proposed SDRDL, an objective function is defined and an iterative optimization algorithm is presented to alternatively optimize the dictionary D and the projection P. In each iteration, SDRDL updates the dictionary D by fixing the projection P and refines the projection P by fixing the dictionary D. After several iterations, the DR projection matrix P and class-specific dictionary D can be obtained together. Therefore, the simultaneously learned P and D will match with each other better. So that more effective pattern classification can be performed by the representation residual. In addition, both the representation residual and the representation coefficients of a query sample will be

discriminative, thus a corresponding classification scheme is presented to exploit such discriminative information. The extensive experiments on image classification tasks showed that SDRDL could achieve competitive performance with those state-of-the-art DL methods.

The rest of this paper is organized as follows. Section 2 briefly introduces the SRC scheme in [42]. Section 3 presents the proposed SDRDL model and its optimization procedure. Section 4 presents the SDRDL based classifier. Section 5 conducts experiments and Section 6 concludes the paper.

2 Sparse representation based classification

Wright et al. [42] proposed the sparse representation based classification (SRC) scheme for robust face recognition (FR). Given K classes of subjects, let $D=[A_1, A_2, \cdots, A_K]$ be the dictionary formed by the set of training samples, where A_i is the subset of training samples from class i. Let y be a query sample. The algorithm of SRC is summarized as follows.

(a) Normalize each training sample in A_i, $i=1,2,\cdots,K$.
(b) Solve l_1-minimization problem: $\hat{x} = \mathrm{argmin}_x \left\{ \|y-Dx\|_2^2 + \gamma\|x\|_1 \right\}$, where γ is scalar constant.
(c) Label a query sample y via: $\mathrm{Label}(y) = \mathrm{argmin}_i\{e_i\}$, where $e_i = \left\|y-A_i\hat{\alpha}^i\right\|_2^2$, $\hat{x} = \left[\hat{\alpha}^1, \hat{\alpha}^2, \cdots, \hat{\alpha}^K\right]^T$ and $\hat{\alpha}^i$ is the coefficient vector associated with class i.

Obviously, the underlying assumption of this scheme is that a query sample can be represented by a weighted linear combination of just those training samples belonging to the same class. Its impressive performance reported in [42] showed that sparse representation is naturally discriminative.

3 Simultaneous dimensionality reduction and dictionary learning (SDRDL)

3.1 Model construction

In most of the previous DL methods introduced in Section 1, the DR and DL processes are performed separately. First, the DR projection matrix is learned from the original training data, then DL is performed to learn a dictionary from the dimensionality reduced training data. Different from most previous DL methods, we propose to learn the DR matrix P and the dictionary D simultaneously for exploiting the discrimination information in the training set more effectively.

Given K classes of subjects, let $A=[A_1, A_2, \cdots, A_K]$ be the set of training samples, where A_i is the subset of training samples from class i. The dimensionality reduced data of training set A can be obtained by PA and it should be represented by the dictionary D with the representation matrix Z, i.e., $PA \approx DZ$. Z can be written as $Z=[Z_1, Z_2, \cdots, Z_K]$, where Z_i is the representation matrix of PA_i over D. Beyond requiring that the dimensionality reduced data PA can be well represented by D (i.e., $PA \approx DZ$), we also require that both of them can cooperate with each other to distinguish the samples in A. Therefore, we propose to simultaneously learn P and

$D=[D_1, D_2, \cdots, D_K]$, where D_i is the class-specific sub-dictionary associated with class i, for performing pattern classification by simultaneous dimensionality reduction and dictionary learning (SDRDL) model:

$$\langle P, D, Z\rangle = \text{argmin}_{P,D,Z}\left\{\sum_{i=1}^{K} r(P, A_i, D, Z_i) + \lambda_1\|Z\|_1 + \lambda_2 h(Z)\right\} \quad s.t. \|d_n\|_2 = 1, \forall n, PP^T = I. \tag{1}$$

where $r(P, A_i, D, Z_i)$ is the projective representation-constrained term; $\|Z\|_1$ is the sparsity penalty; $h(Z)$ is coefficients diversity term imposed on the coefficient matrix Z; λ_1, λ_2, and γ are scalar parameters. Each atom d_n of D is constrained to have a unit l_2-norm to avoid that D has arbitrarily large l_2-norm, resulting in trivial solutions of the coefficient matrix Z. In the following section, we will discuss the terms $r(P, A_i, D, Z_i)$ and $h(Z)$ in details.

3.1.1 Projective representation-constrained term $r(P, A_i, D, Z_i)$

The dimensionality reduced data PA_i of the original data A_i should be represented by D with Z_i, i.e., $PA_i \approx DZ_i$. Z_i can be written as $Z_i=[Z_i^1; \cdots; Z_i^j; \cdots; Z_i^K]$, where Z_i^j is the representation coefficients of PA_i over D_j. Denote by $R_k = D_k Z_i^k$ the representation of D_k to PA_i. There is:

$$PA_i \approx DZ_i = D_1 Z_i^1 + \cdots + D_i Z_i^i + \cdots + D_k Z_i^k = R_1 + \cdots + R_i + \cdots + R_k \tag{2}$$

where $R_i = D_i Z_i^i$. Since D_i is associated with the i^{th} class, it is naturally expected that PA_i could be well represented by D_i but not by $D_j, j \neq i$. This implies that there are some significant coefficients in Z_i^i such that $\|PA_i - D_i Z_i^i\|_F^2$ is small, while some coefficients in Z_i^j such that $\|PA_i - D_j Z_i^j\|_F^2$ is big. Making $\|PA_i - D_j Z_i^j\|_F^2$ big can be attained, to some extent, by making Z_i^j having some very small coefficients such that $\|D_j Z_i^j\|_F^2$ is small. Furthermore, since PA_i is the dimensionality reduced data of the original data A_i, it is also expected that A_i can be well reconstructed from the projected subspace by P. This can be accomplished by minimizing $\|A_i - P^T PA_i\|_F^2$, which is the amount of energy discarded by the DR matrix P or the difference between low-dimensional approximations and the original training set. Therefore, in this work, we define projective representation-constrained term $r(P, A_i, D, Z_i)$ as:

$$r(P, A_i, D, Z_i) = \|PA_i - DZ_i\|_F^2 + \left\|PA_i - D_i Z_i^i\right\|_F^2 + \sum_{j=1, j\neq i}^{K} \left\|D_j Z_i^j\right\|_F^2 + \gamma\left\|A_i - P^T PA_i\right\|_F^2 \tag{3}$$

By using these four terms in Eq. (3), A_i can be well reconstructed by $P^T PA_i$, also P and D can be better fit with each other. So that D_i will have not only the minimal but also very small representation residual for PA_i, while other class-specific sub-dictionary $D_j, j \neq i$ will have big representation residuals of PA_i.

3.1.2 Coefficients diversity term $h(Z)$

Given a query sample, Wright et al. proposed that its sparse representation could be found by SRC scheme and the largest coefficients in the coefficients vector recovered by SRC are associated with the training samples, which have the same class label as the query sample [42]. It implies that the query sample can be approximated by a weighted linear combination of its own training samples with these largest coefficients. Likewise, in our proposed SDRDL

model, it is expected that the largest coefficients in Z_i (or Z_j) are associated with D_i (or D_j), as illustrated in Fig. 1. From this figure, it can also be found that when A_i and A_j belong to the same class, $\|Z_j^T Z_i\|_F^2$ is big, while when A_i and A_j belong to different classes, $\|Z_j^T Z_i\|_F^2$ is small. Actually, $\|Z_j^T Z_i\|_F^2$ reflects the relative similarity of samples. Thus, coefficients diversity term $h(Z)$ can be defined as:

$$h(Z) = \sum_{j \neq i} \left\| Z_j^T Z_i \right\|_F^2 \tag{4}$$

When A_i and A_j belong to different classes (i.e., $i \neq j$), minimizing the coefficients diversity term $h(Z)$ encourages that the largest coefficients in Z_i and Z_j are associated with the corresponding different sub-dictionary (i.e., D_i and D_j, respectively). Therefore, the discriminative ability of dictionary D can be further promoted.

3.1.3 SDRDL model

By incorporating Eqs. (3) and (4) into Eq. (1), SDRDL model can be formulated as:

$$\langle P, D, Z \rangle = \operatorname{argmin}_{P,\, D, Z} \left\{ \sum_{i=1}^{K} \left(\|PA_i - DZ_i\|_F^2 + \left\|PA_i - D_i Z_i^i\right\|_F^2 + \sum_{j=1, j \neq i}^{K} \left\|D_j Z_i^j\right\|_F^2 + \gamma \left\|A_i - P^T PA_i\right\|_F^2 \right) + \lambda_1 \|Z_1\| + \lambda_2 \sum_{j \neq i} \left\| Z_j^T Z_i \right\|_F^2 \right\} s.t. \ \|d_n\|_2 = 1, \ \forall n, \ PP^T = I. \tag{5}$$

The objective of SDRDL is to simultaneously learn the DR matrix P as well as the class-specific dictionary D. In a subspace determined by P, each sub-dictionary D_i will have small representation residuals to the samples from class i but have big representation residuals to other classes. Besides, the representation coefficient vectors of samples from one class will be similar to each other but dissimilar to samples from other classes. Ideally, if P and D could be well optimized, they will be proper for classification task.

3.2 Optimization

Obviously, Eq. (5) is a single-objective optimization problem and its objective function is non-convex for $\langle P, D, Z \rangle$. As the works [19, 33, 49, 50, 52] have done when trying to solve similar

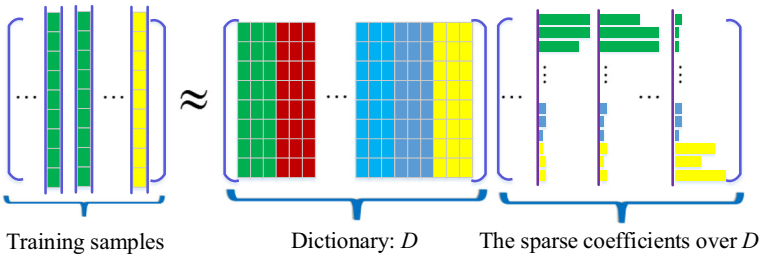

Fig. 1 Sparse representation of training samples over dictionary D. Training samples with different colors belong to different classes. And atoms with different colors in dictionary D have different class labels

optimization problems, here we divide the whole optimization into two sub-problems: updating D and Z by fixing P; and updating P by fixing D and Z. These two sub-problems are solved alternatively and iteratively for the desired dimensionality reduction projection matrix P and dictionary D.

3.2.1 Update $\boldsymbol{D}$ and $\boldsymbol{Z}$

Suppose that P is fixed, D and Z are updated. When P is fixed, the objective function in Eq. (5) reduces to:

$$\langle D,Z\rangle = \mathrm{argmin}_{D,Z}\left\{\sum_{i=1}^{K}\left(\|PA_i - DZ_i\|_F^2 + \left\|PA_i - D_iZ_i^i\right\|_F^2 + \sum_{j=1,j\neq i}^{K}\left\|D_jZ_i^j\right\|_F^2\right) + \lambda_1 Z_1 + \lambda_2\sum_{j\neq i}\left\|Z_j^TZ_i\right\|_F^2\right\} s.t. \|d_n\|_2 = 1 \tag{6}$$

Let $B_i = PA_i$, we rewrite Eq. (6) as:

$$\langle D,Z\rangle = \mathrm{argmin}_{D,Z}\left\{\sum_{i=1}^{K}\left(\|B_i - DZ_i\|_F^2 + \left\|B_i - D_iZ_i^i\right\|_F^2 + \sum_{j=1,j\neq i}^{K}\left\|D_jZ_i^j\right\|_F^2\right) + \lambda_1 Z_1 + \lambda_2\sum_{j\neq i}\left\|Z_j^TZ_i\right\|_F^2\right\} s.t. \|d_n\|_2 = 1 \tag{7}$$

Obviously, D and Z can be solved alternatively and iteratively. When D is fixed, the objective function in Eq. (7) can be reduced to update $Z = [Z_1, Z_2, \cdots, Z_K]$. Z_i can be computed class by class. Thus the objective function in Eq. (7) can be further reduced to:

$$\min_{Z_i}\left\{\|B_i - DZ_i\|_F^2 + \left\|B_i - D_iZ_i^i\right\|_F^2 + \sum_{j=1,j\neq i}^{K}\left\|D_jZ_i^j\right\|_F^2 + \lambda_1\|Z_i\|_1 + \lambda_2\sum_{j=1,j\neq i}^{K}\left\|Z_j^TZ_i\right\|_F^2\right\} \tag{8}$$

To prevent Z_j from having arbitrarily large l_2-norm, we normalize each column of Z_j in Eq. (8) to a unit l_2-norm. Thus, Z_i can be computed by the following objective function:

$$\min_{Z_i}\left\{\|B_i - DZ_i\|_F^2 + \left\|B_i - D_iZ_i^i\right\|_F^2 + \sum_{j=1,j\neq i}^{K}\left\|D_jZ_i^j\right\|_F^2 + \lambda_1\|Z_i\|_1 + \lambda_2\sum_{j=1,j\neq i}^{K}\left\|\tilde{Z}_j^TZ_i\right\|_F^2\right\} \tag{9}$$

where, $\tilde{Z}_j = \left[\tilde{z}_{j,1}, \tilde{z}_{j,2}, \cdots, \tilde{z}_{j,n_j}\right]$ denotes the normalized Z_j and $\tilde{z}_{j,i} = z_{j,i}/\left\|z_{j,i}\right\|_2$, $i = 1, 2, \cdots, n_j$. We rewrite Eq. (9) as:

$$\min_{Z_i}\left\{\varphi_i(Z_i) + \lambda_1\|Z_i\|_1\right\} \tag{10}$$

where, $\varphi_i(Z_i) = \|B_i - DZ_i\|_F^2 + \left\|B_i - D_iZ_i^i\right\|_F^2 + \sum_{j=1,j\neq i}^{K}\left\|D_jZ_i^j\right\|_F^2 + \lambda_2\sum_{j=1,j\neq i}^{K}\left\|\tilde{Z}_j^TZ_i\right\|_F^2$. It can be proved that $\varphi_i(Z_i)$ is convex with Lipschitz continuous gradient (please refer to Appendix for the proof). Therefore, in this work we can adopt a new fast iterative shrinkage-thresholding algorithm (FISTA) [2] to solve Eq. (10), as described in Table 1.

Table 1 Learning sparse code Z_i

Algorithm of Learning Sparse Codes Z_i
1. **Input:** DR projection matrix P; a training subset A_i from class i; initial Z^0; dictionary D; the parameters $\rho, \tau > 0$.
2. **Initialization:** $\hat{Z}_i^{(1)} \leftarrow Z^0$ and $t \leftarrow 1$;
3. **While** not converge or the maximal iteration step is not reached **do**
$t \leftarrow t+1$;
$u^{(t-1)} \leftarrow \hat{Z}_i^{(t-1)} - {}^1/_{2\rho}\nabla\varphi_i(\hat{Z}_i^{(t-1)})$, where $\nabla\varphi_i(\hat{Z}_i^{(t-1)})$ is the derivative of $\varphi_i(\hat{Z}_i^{(t-1)})$ w.r.t. $\hat{Z}_i^{(t-1)}$;
$\hat{Z}_i^{(t)} \leftarrow soft(u^{(t-1)}, {}^\tau/_\rho)$, where $soft(u, {}^\tau/_\rho)$ is defined by [43]:
$\left[soft(u, {}^\tau/_\rho)\right]_j = \begin{cases} 0 & \lvert u_j\rvert \le {}^\tau/_\rho \\ u_j - sign(u_j){}^\tau/_\rho & otherwise \end{cases}$;
4. **Output:** $\hat{Z}_i = \hat{Z}_i^{(t)}$

When Z is fixed, the objective function in Eq. (7) can be reduced to compute $D=[D_1, D_2, \cdots, D_K]$. $D_i = \left[d_1, d_2, \cdots, d_{p_i}\right]$ is also updated class by class. Thus the objective function in Eq. (7) can be reduced as:

$$\min_{D_i}\left\{\left\|\overline{B}-D_iZ^i\right\|_F^2 + \left\|B_i-D_iZ_i^i\right\|_F^2 + \sum\nolimits_{j=1,j\neq i}^{K}\left\|D_iZ_j^i\right\|_F^2\right\} s.t. \|d_l\|_2 = 1, l = 1, 2, \cdots, p_i \quad (11)$$

where $\overline{B} = B - \sum_{j=1,j\neq i}^{K} D_jZ^j$ and $B=[B_1, B_2, \cdots, B_K]$; Z^i represent the coefficient matrix of B over D_i. Equation (11) can be rewritten as the following form [49, 50]:

$$\min_{D_i}\left\|\overline{B}_i - D_iX_i\right\|_F^2 s.t. \|d_l\|_2 = 1, l = 1, 2, \cdots, p_i \quad (12)$$

where $\overline{B}_i = \left[\overline{B}\ B_i\ 0\cdots00\cdots0\right]$, $X_i=[Z^i\ Z_i^i\ Z_1^i\cdots Z_{i-1}^i\ Z_{i+1}^i\cdots Z_K^i]$. Equation (12) can be efficiently solved by updating each dictionary atom one by one via the algorithm [23, 49, 50] as presented in Table 2.

Table 2 Learning dictionary D_i

Algorithm of Learning Sparse Codes D_i
1. **Input:** DR projection matrix P; a training set A; initial dictionary D_i; coefficients X_i.
2. Let $B = PA$, $X_i = \left[x_1; x_2; \cdots; x_{p_i}\right]$ and $D_i = \left[d_1, d_2, \cdots, d_{p_i}\right]$, where x_j is the j^{th} row vector of X_i and d_j is the j^{th} column vector of D_i, $j = 1, 2, \cdots, p_i$.
3. **For** $j = 1$ to p_i do
Fix all $d_l, l \neq j$ update d_j. Let $Y = \overline{\overline{B}}_i - \sum_{l\neq j} d_lx_l$. The minimization of Eq. (11) becomes:
$\min_{d_j}\left\|Y - d_jx_j\right\|_F^2 s.t. \left\|d_j\right\|_2 = 1;$
By solving this objective function [17, 33, 34], we could get the solution $d_j = Yx_j^T/\left\|Yx_j^T\right\|_2$.
4. **Output:** the updated version of D_i

3.2.2 Update $\boldsymbol{P}$

Suppose that D and Z are fixed, P is updated. When D and Z are fixed, the object function in Eq. (5) can be reduced to:

$$P = \mathrm{argmin}_P \left\{ \sum_{i=1}^{K} \|PA_i - DZ_i\|_F^2 + \left\|PA_i - DZ_i^i\right\|_F^2 + \gamma \left\|A_i - P^T PA_i\right\|_F^2 \right\} s.t. PP^T = I. \quad (13)$$

Let $W_i = DZ_i$ and $W_i^i = DZ_i^i$, Eq. (13) can be re-formulated as:

$$\begin{aligned} P &= \mathrm{argmin}_P \left\{ \sum_{i=1}^{K} \|PA_i - W_i\|_F^2 + \left\|PA_i - W_i^i\right\|_F^2 + \gamma \left\|A_i - P^T PA_i\right\|_F^2 \right\} \\ &= \mathrm{argmin}_P \sum_{i=1}^{K} \left\{ tr\left(PQ_i(P)P^T\right) + tr\left(PQ_i^i(P)P^T\right) + \gamma tr\left(A_i^T A_i - PA_i A_i^T P^T\right) \right\} \\ &\quad s.t.\ PP^T = I. \end{aligned} \quad (14)$$

where $Q_i(P) = (A_i - P^T W_i)(A_i - P^T W_i)^T$ and $Q_i^i(P) = (A_i - P^T W_i^i)(A_i - P^T W_i^i)^T$. Since the term $A_i^T A_i$ has no effect on the solution of P, the objection function in Eq. (14) further reduces to:

$$\begin{aligned} P &= \mathrm{argmin}_P \sum_{i=1}^{K} \left\{ tr\left(PQ_i(P)P^T\right) + tr\left(PQ_i^i(P)P^T\right) - \gamma tr\left(PA_i A_i^T P^T\right) \right\} \\ &= \mathrm{argmin}_P tr \left\{ P \sum_{i=1}^{K} \left(Q_i(P) + Q_i^i(P) - \gamma A_i A_i^T\right) P^T \right\} s.t. PP^T = I. \end{aligned} \quad (15)$$

The above minimization can be solved iteratively. In the current iteration t, we use $Q_i(P^{(t-1)})$ and $Q_i^i(P^{(t-1)})$ to approximate $Q_i(P^{(t)})$ and $Q_i^i(P^{(t)})$ in Eq. (15), where $P^{(t-1)}$ is the projection matrix obtained in iteration $t-1$. By applying the Eigen Value Decomposition (EVD) technique, we have:

$$[U, M, V] = EVD\left(\sum_{i=1}^{K} \left(Q_i(P) + Q_i^i(P) - \gamma A_i A_i^T\right) \right) \quad (16)$$

where M is diagonal matrix formed by the eigenvalues of $\sum_{i=1}^{K} (Q_i(P) + Q_i^i(P) - \gamma A_i A_i^T)$. Then we can update P as the m eigenvectors in U associated with the first m smallest eigenvalues of M, i.e. $P^{(t)} = U(1:m, :)$. However, this way makes the update of P too sharp and the optimization of the whole system in Eq. (5) unstable. Therefore, we choose to update P gradually in our implementation and thus have the following form for updating P in each iteration:

$$P^{(t)} = P^{(t-1)} + c\left(U(1:m, :) - P^{(t-1)}\right) \quad (17)$$

where c is a small positive constant and used to control the update of P in iterations. The algorithm of updating P is presented in Table 3.

Table 3 Learning dimensionality reduction projection matrix P

Algorithm of learning the dimensionality reduction projection matrix P
1. Input: initial P^0; a training set A; dictionary D; coefficients matrix Z
2. Let $W_i = DZ_i$ and $W_i^i = DZ_i^i$, $i = 1,2,\cdots,K$.
3. Initialization: $\hat{P}^{(1)} \leftarrow P^0$ and $t \leftarrow 1$;
4. While not converge or the maximal iteration step is not reached **do**
$t \leftarrow t+1$;
$Q_i(\hat{P}^{(t)}) \leftarrow \left(A_i - \hat{P}^{(t-1)^T} W_i\right)\left(A_i - \hat{P}^{(t-1)^T} W_i\right)^T, i = 1,2,\cdots,K;$
$Q_i^i(\hat{P}^{(t)}) \leftarrow \left(A_i - \hat{P}^{(t-1)^T} W_i^i\right)\left(A_i - \hat{P}^{(t-1)^T} W_i^i\right)^T, i = 1,2,\cdots,K;$
$\hat{P}^{(t)} \leftarrow \hat{P}^{(t-1)} + c(U(1{:}m,:) - \hat{P}^{(t-1)})$, where U is defined as:
$[U,M,V] = EVD(\sum_{i=1}^{K}(Q_i(P) + Q_i^i(P) - \gamma A_i A_i^T))$;
5. Output: $\hat{P} = \hat{P}^{(t)}$.

3.2.3 Algorithm of SDRDL

The complete SDRDL algorithm is summarized in Table 4. The proposed SDRDL model in Eq. (5) is jointly non-convex to $\langle P, D, Z\rangle$, and thus the optimization algorithm presented in Table 4 can at most attain a local minimum. When SDRDL updates D and Z, the objection function in Eq. (6) is convex to each of $\langle D, Z\rangle$ by fixing the other, and the proposed SDRDL algorithm will lead to a local minimum of this sub-problem. However, when SDRDL updates P, $Q_i(P^{(t-1)})$ and $Q_i^i(P^{(t-1)})$ are used to approximate $Q_i(P^{(t)})$ and $Q_i^i(P^{(t)})$ in Eq. (15), thereby the obtained solution is only an approximation to the local minimum of the objection function in Eq. (15). Overall, the proposed SDRDL algorithm cannot converge in theory, but by experience can have a stable solution.

To illustrate the minimization process of SDRDL, we use the Extended Yale B database [13] and AR database [26] as examples. The reduced dimensionality of the face images is set to be 300. The curves of the objective function in Eq. (5) vs. the iteration number are plotted in Fig. 2a and b for these two databases, respectively. It can be seen that after 15 iterations, the

Table 4 The algorithm of simultaneous dimensionality reduction and dictionary learning

Simultaneous dimensionality reduction and dictionary learning (SDRDL)
1. Initialize D.
Initialize the P as the PCA transformation matrix of the training data A.
Initialize the atoms of D_i as the eigenvectors of PA_i;
2. While convergence or the maximal iteration step is not reached **do**
$t \leftarrow t+1$;
Fix P and D, update $Z_i, i = 1,2,\cdots,K$, one by one with the algorithm in **Table 1**;
Fix P and Z, update $D_i, i = 1,2,\cdots,K$, one by one with the algorithm in **Table 2**;
Fix D and Z, update P with the algorithm in **Table 3**;
3. Output: P, Z and D.

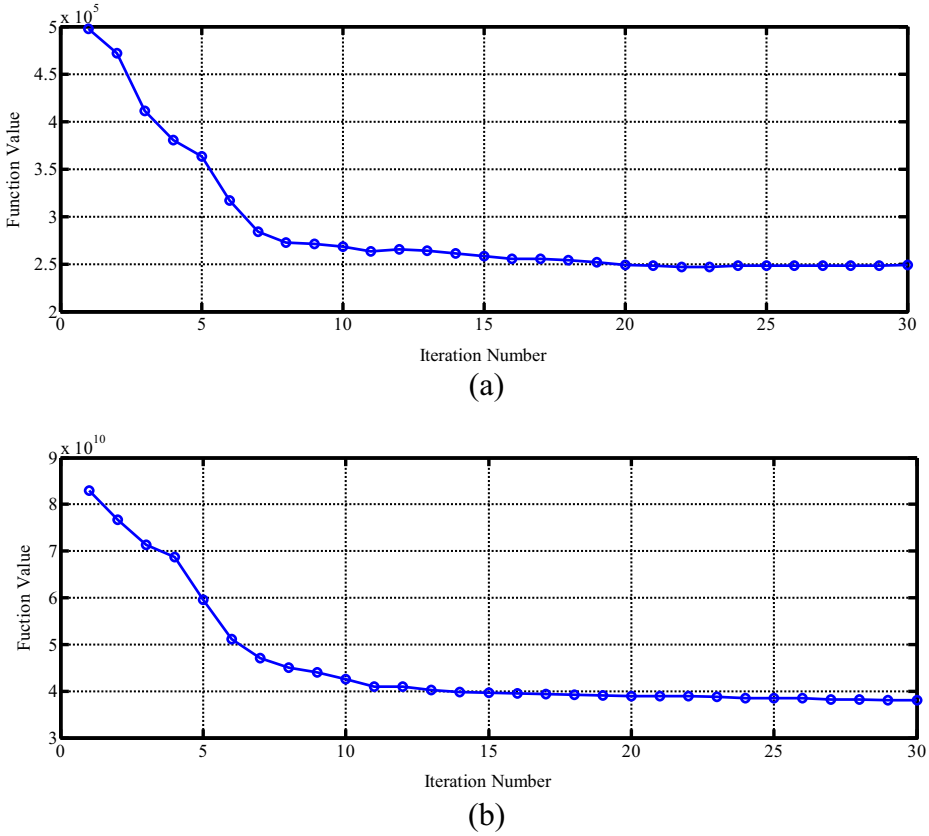

Fig. 2 Convergence curves of SDRDL algorithm on **a** Extended Yale B database **b** AR database

value of the objective function has small variation and becomes stable on these two databases. Our experiment results also indicate that when the minimization stops with more or less than 15 iterations, the learned projection P and dictionary D by SDRDL will lead to almost the same classification accuracy. Therefore, we choose to set the maximal iteration number as 15 and it works well in our experiments.

4 Classification scheme

Once we obtain the DR projection matrix P and the class-specific dictionary D, the lower dimensional feature of the query sample y can be computed by Py and then it can be coded over the dictionary D. Here, the sparse representation model with l_1-norm is adopted for coding:

$$\hat{\alpha} = \text{argmin}_{\alpha}\left\{\|Py - D\alpha\|_2^2 + \lambda\|\alpha\|_1\right\} \tag{18}$$

where λ is constant.

Denote by $\hat{\alpha} = \left[\hat{\alpha}^1, \hat{\alpha}^2, \cdots, \hat{\alpha}^K\right]^T$, where $\hat{\alpha}^i$ is the coefficient sub-vector associated with sub-dictionary D_i. Once $\hat{\alpha}$ is computed, the reconstruction residual of each D_i can be used to classify an input query sample, as that in SRC [42]. On the other hand, the normalized coefficient matrix of each class, denoted by $\tilde{Z}_j$, is also learned in the proposed SDRDL algorithm and thus the dissimilarity between $\tilde{Z}_j$ and $\hat{\alpha}$ is also discriminative. Therefore, the metric for classification can be defined as:

$$e_i = \left\| Py - D_i \hat{\alpha}^i \right\|_2^2 + w \sum\nolimits_{j \neq i} \tilde{Z}_j^T \hat{\alpha}_F^2 / n_j \tag{19}$$

where w is a preset weight to balance the contribution of the two terms for classification; n_j is the number of training samples from class j. When the number of training samples from each class is the same, Eq. (19) can be rewritten as:

$$e_i = \left\| Py - D_i \hat{\alpha}^i \right\|_2^2 + \delta \sum\nolimits_{j \neq i} \left\| \tilde{Z}_j^T \hat{\alpha} \right\|_F^2 \tag{20}$$

where $\delta = w/n_j$. The classification rule is simply set as identity$(y) = \text{argmin}_i\{e_i\}$.

5 Experimental results

We verify the performance of SDRDL on applications such as face recognition and action classification. Section 5.1 discusses parameter selection; Section 5.2 conducts experiments on face recognition; Section 5.3 perform experiments on action classification. In Tables 5, 6, and 7, the highest classification rates are highlighted in boldface.

5.1 Parameter selection

In SDRDL, it is very important to set the number of atoms in D_i, denoted by $p_i, i = 1, 2, \cdots, K$. Usually, each p_i is set to be equal. To evaluate the effect of p_i on the performance of SDRDL, we conduct face recognition experiment on the Extended Yale B database [13], which consists of

Table 5 The face recognition rates (%) of competing methods on the Extended Yale B database

Methods	Accuracy
NNC	61.7
SVM	88.8
SRC [42]	90.0
DKSVD [52]	75.3
LC-KSVD [19]	90.6
DLSI [33]	85.0
DLSI* [33]	89.0
FDDL [49, 50]	91.9
SDRDL	**96.7**

Table 6 The face recognition rates (%) of competing methods on the AR database

Methods	Accuracy
NNC	71.4
SVM	87.1
SRC [42]	88.8
DKSVD [52]	85.4
LC-KSVD [19]	89.7
DLSI [33]	73.7
DLSI* [33]	89.8
FDDL [49, 50]	92.0
SDRDL	**93.3**

2414 frontal-face images from 38 individuals. For each individual, 20 images are picked randomly for training; the remaining images (about 44 images per individual) are used for testing. In addition, SRC is used as the baseline method in this experiment. Since SRC uses the original training samples as dictionary, we pick randomly p_i training samples as the dictionary atoms and conduct ten times the experiment to calculate the average classification accuracy. Figure 3 shows the classification accuracies of SDRDL and SRC vs. the number of dictionary atoms.

It can be seen that SDRDL achieves 4.0 % improvement at least over SRC. Even with $p_i = 8$, SDRDL yet obtains higher classification accuracy than SRC with $p_i = 20$. In addition, when p_i decreases from 20 to 8, the classification accuracy of SDRDL drops by 2.5 %, while the classification accuracy of SRC drops by 4.2 %. This demonstrates that SDRDL is effective to learn a compact and representative dictionary, by which the computation complexity can be reduced and the classification accuracy can be improved.

There are five parameters which need to be tuned in the proposed SDRDL model, three in the DL model (λ_1, λ_2 and γ) and two in the classification scheme (λ and w (or δ)). In all the experiments, if no specific instructions, the tuning parameters in SDRDL are evaluated by fivefold cross validation.

Table 7 The accuracies (%) of competing methods on UCF sports action database

Methods	Accuracy
Qiu et al. [32]	83.6
Sadanand et al. [36]	90.7
Yao et al. [51]	86.6
SRC [42]	92.9
KSVD [1]	86.8
DKSVD [52]	88.1
LC-KSVD [19]	91.2
DLSI [33]	92.1
COPAR [20]	90.7
JDL [54]	90.0
FDDL [49, 50]	94.3
SDRDL	**94.5**

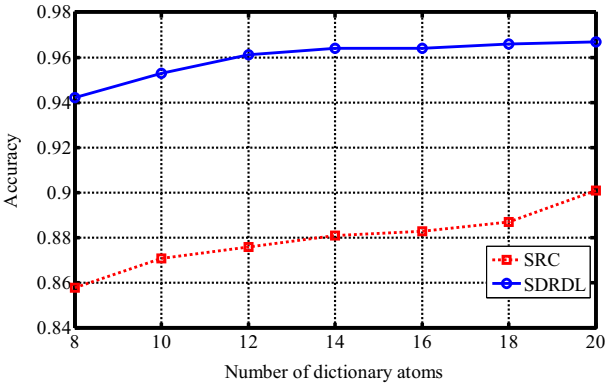

Fig. 3 The Accuracy of SDRDL and SRC vs. the number of dictionary atoms

5.2 Face recognition

The proposed SDRDL method is applied to FR on the Extended Yale B [13] and AR [26] databases and compared with several latest DL based FR methods including discriminative KSVD (DKSVD) [52], label consistent KSVD (LC-KSVD) [19], dictionary learning with structure incoherence (DLSI) [33] and Fisher discrimination dictionary learning (FDDL) [49, 50]. In addition, SDRDL is also compared with SRC [42] and two general classifiers, nearest neighbor classifier (NNC) and linear support vector machine (SVM). Note that the original DLSI method represents the query sample class by class. For a fair comparison, we extend the original DLSI by representing the query sample over the whole dictionary and then use the representation residual for pattern classification (denoted by DLSI* respectively). The number of dictionary atoms in SDRDL is set as the number of training samples in default. Each face image is reduced to dimension of 300 in all FR experiments. Since the number of training samples from each class is equal in all of face databases, Eq. (20) is adopted to perform classification for all FR experiments. The parameters of SDRDL chosen by cross-validation are $\lambda_1 = 0.005$, $\lambda_2 = 0.07$, $\gamma = 0.9$, $\lambda = 0.005$ and $\delta = 0.8$.

5.2.1 Extended Yale B database

The Extended Yale B database [13] consists of 2414 frontal face images from 38 individuals taken under varying illumination conditions (see Fig. 4 for example samples). Each individual has 64 images and we randomly selected 20 images for training and the remaining images for testing. The face image is normalized to 54×48.

Fig. 4 Some samples from the Extended Yale B database

The results of SDDRDL, NNC, SRC, SVM, DKSVD, LC-KSVD, DLSI and FDDL are listed in Table 5. It can be seen that the proposed SDRDL achieves the highest classification accuracy among the competing methods. Particularly, SDRDL achieves 4.8 % higher classification accuracy than FDDL, which achieves the second best performance in the experiment. The reason may be that, FDDL performs dimensionality separately from the discriminative dictionary learning process. SDRDL and FDDL outperform DKSVD and LC-KSVD, which only use representation coefficients to perform classification. DLSI* has 4 % higher classification accuracy than DLSI. This can be explained that representing the query image on the whole dictionary is more reasonable for FR tasks.

5.2.2 AR database

The AR database [26] consists of over 4000 images of 126 individuals. For each individual, 26 pictures are collected from two separated sessions. Following [49, 50], we chose a subset consisting of 50 male individuals and 50 female individuals for the standard evaluation procedure. For each individual, we select 7 images with illumination and expression changes from Session 1 for training and 7 images with the same condition from Session 2 for testing. The face image is of size 60×43. Figure 5 shows sample images from the AR database.

The classification accuracy of SDRDL and other competing methods including NNC, SRC, SVM, DKSVD, LC-KSVD, DLSI and FDDL are shown in Table 6. Again, it can be seen that the proposed SDRDL outperforms FDDL and achieves the highest classification rate. It is because that by coupling the dimensionality reduction and dictionary learning processes, SDRDL can exploit the discriminative information in training set more effectively for FR task. Being consistent with the results on Extended Yale B database, FDDL and SDRDL achieve higher classification rate than DKSVD and LC-KSVD. Also, DLSI* has much better result than DLSI. This further demonstrate that representing the query image on the whole dictionary is more reasonable for FR tasks.

5.3 Action recognition

Finally, we evaluated SDRDL on the UCF sports action dataset [35] for action classification experiment. The UCF sports action dataset [35] includes 140 videos that are collected from various broadcast sports channels (e.g., BBC and ESPN). These videos cover a wide range of scenarios and viewpoints. The actions of these videos contain 10

Fig. 5 Some samples from the AR database

Fig. 6 Some video frames from the UCF sports action dataset

sport action classes: driving, golfing, kicking, lifting, horse riding, running, skateboarding, swinging-(prommel horse and floor), swinging-(high bar) and walking. Some example images of this dataset are shown in Fig. 6. The action bank features of 140 videos provided by [36] are adopted in the experiment.

As the experiment setting in [32] and [19], we evaluated SDRDL via five-fold cross validation. The number of atoms in per sub-dictionary is set as the number of training samples in SDRDL with $\lambda_1 = 0.005$, $\lambda_2 = 0.04$, $\gamma = 0.9$ and $\lambda = 0.005$. Because the number of training samples from each class is not equal in this experiment, Eq. (19) is adopted to perform classification. Here w is set to 2.0 and the reduced dimension of the action bank feature [36] is set to be 100. SDRDL is compared with SVM, SRC [42], KSVD [1], DKSVD [52], LC-KSVD [19], DLSI [33], FDDL [49, 50], dictionary learning with commonality and particularity (COPAR) [20] and joint dictionary learning (JDL) [54]. The performance of some specific methods for action recognition including Yao et al. [51] and Qiu et al. [32] are also listed. The classification accuracies are shown in Table 7. It can be observed that SDRDL obtains the best performance. Following SDRDL, FDDL shows the second best performance. With the action bank feature, all the dictionary learning (DL) based methods obtain the classification accuracy over 90 % except KSVD and DKSVD.

6 Conclusion

In this paper we proposed a simultaneous dimensionality reduction and dictionary learning (SDRDL) scheme for image classification. Unlike many DL methods which perform dimensionality reduction and dictionary learning independently, SDRDL formulates DR and DL procedures into a unified framework to exploit the discriminative information more effectively in the training data. In classification, both the representation residual and the representation coefficients were considered. The experimental results on benchmark image databases demonstrated that the proposed SDRDL method surpasses many state-of-the-arts methods.

Acknowledgments This work was supported by National Instrument Development Special Program of China under the grants 2013YQ03065101, 2013YQ03065105, Ministry of Science and Technology of China under National Basic Research Project under the grants 2010CB731803, and by National Natural Science Foundation of China under the grants 61221003, 61290322, 61174127, 61273181, 60934003, 61290322, 61503243 and U1405251, the Program of New Century Talents in University of China under the grant NCET-13-0358, the Science and Technology Commission of Shanghai Municipal, China under the grant 13QA1401900, Postdoctoral Science Foundation of China under the grants 2014 M551406.

Appendix

$\varphi_i(Z_i)$ is convex and continuously differentiable with Lipschitz continuous gradient $L(\varphi_i)$:

$$\|\nabla\varphi_i(x)-\nabla\varphi_i(y)\|\le L(\varphi_i)\|x-y\|, \forall x, y\in R^n. \tag{21}$$

where $\|\cdot\|$ denotes the standard Euclidean norm and $L(\varphi_i)>0$ is the Lipschitz constant of $\nabla\varphi_i$.

In Eq. (10)

$$\varphi_i(Z_i) = \|B_i-DZ_i\|_F^2 + \left\|B_i-D_iZ_i^i\right\|_F^2 + \sum\nolimits_{j\neq i}\left\|D_jZ_i^j\right\|_F^2 + \lambda_2\sum\nolimits_{j\neq i}\left\|\tilde{Z}_j^TZ_i\right\|_F^2 \tag{22}$$

Let $Z_i^i = P^iZ_i$ and $Z_i^j = P^jZ_i$ where $P^i(P^j)$ are projection matrixes which keeps components of $Z_i(Z_j)$ associated with $D_i(D_j)$ unchanged but sets other components to be zero. Hence, we can rewrite Eq. (22) as:

$$\varphi_i(Z_i) = \|B_i-DZ_i\|_F^2 + \left\|B_i-DP^iZ_i\right\|_F^2 + \sum\nolimits_{j\neq i}\left\|DP^jZ_i\right\|_F^2 + \lambda_2\sum\nolimits_{j\neq i}\left\|\tilde{Z}_j^TZ_i\right\|_F^2 \tag{23}$$

Let $DP^i = D^i$ and $DP^j = D^j$. Equation (23) equals to:

$$\varphi_i(Z_i) = \|B_i-DZ_i\|_F^2 + \left\|B_i-D^iZ_i\right\|_F^2 + \sum\nolimits_{j\neq i}\left\|D^jZ_i\right\|_F^2 + \lambda_2\sum\nolimits_{j\neq i}\left\|\tilde{Z}_j^TZ_i\right\|_F^2 \tag{24}$$

The stacking operator introduced in [30] can be used to write B_i and Z_i as a column vector. We form $\Psi_i = \left[b_{i,1}, b_{i,2}, \cdots, b_{i,n_i}\right]^T$, $\chi_i = \left[z_{i,1}, z_{i,2}, \cdots, z_{i,n_i}\right]^T$ where $a_{i,i}, z_{i,i}\in R^{m\times 1}$ and thus $\Psi_i, \chi_i\in R^{(m\cdot n_i)\times 1}$. Hence, Eq. (24) can be rewrite as:

$$\begin{aligned}\varphi_i(\chi_i) = \|\Psi_i-\mathrm{diag}(D)\chi_i\|_2^2 + \left\|\Psi_i-\mathrm{diag}\left(D^i\right)\chi_i\right\|_2^2 + \sum\nolimits_{j\neq i}\left\|\mathrm{diag}\left(D^j\right)\chi_i\right\|_2^2 + \\ \lambda_2\sum\nolimits_{j\neq i}\left\|\mathrm{diag}\left(\tilde{\chi}_j^T\right)\chi_i\right\|_2^2\end{aligned} \tag{25}$$

where diag(T) is a block diagonal matrix with each block on the diagonal being matrix T. And also $\varphi_i(\chi_i)$ equals to:

$$\begin{aligned}2\Psi_i^T\Psi_i - 2\Psi_i^T\left(\mathrm{diag}(D) + \mathrm{diag}\left(D^i\right)\right)\chi_i + \chi_i^T\left(\mathrm{diag}\left(D^TD\right) + \mathrm{diag}\left(D^{iT}D^i\right) + \right. \\ \left.\mathrm{diag}\left(\sum\nolimits_{j\neq i}D^{jT}D^j\right) + \lambda_2\mathrm{diag}\left(\sum\nolimits_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\right)\right)\chi_i\end{aligned} \tag{26}$$

The convexity of $\varphi_i(\chi_i)$ depends on its Hessian matrix $\nabla^2\varphi_i(\chi_i)$ is whether positive semi-definite or not [5]. We could write the Hessian matrix of $\varphi_i(\chi_i)$ as:

$$\begin{aligned}\nabla^2\varphi_i(\chi_i) = 2\mathrm{diag}\left(D^TD\right) + 2\mathrm{diag}\left(D^{iT}D^i\right) + 2\mathrm{diag}\left(\sum\nolimits_{j\neq i}D^{jT}D^j\right) + \\ 2\lambda_2\mathrm{diag}\left(\sum\nolimits_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\right)\end{aligned} \tag{27}$$

Since diag(D^TD), diag($D^{iT}D^i$), diag($\sum_{j\neq i}D^{jT}D^j$) and diag $\left(\sum_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\right)$ are all Hermite matrix, they are all positive semi-definite. Therefore, Hessian matrix $\nabla^2\varphi_i(\chi_i)$ is positive semi-definite. Based on this, we claim that $\boldsymbol{\varphi_i(\chi_i)}$ is a convex function.

Via Eq. (26), we have:

$$\nabla\varphi_i(\chi_i) = -2\Psi_i^T\left(\mathrm{diag}(D) + \mathrm{diag}\left(D^i\right)\right) + 2\Big(\mathrm{diag}\left(D^TD\right) + \mathrm{diag}\Big(D^{i^T}D^i\Big) + \mathrm{diag}\Big(\sum\nolimits_{j\neq i} D^{j^T}D^j\Big) + \lambda_2\mathrm{diag}\Big(\sum\nolimits_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\Big)\Big) \quad \chi_i \tag{28}$$

From Eq. (28), we can easy see that $\nabla\varphi_i(\chi_i)$ is continuously differentiable to χ_i. And via Eq. (28), we have:

$$\nabla\varphi_i(x) - \nabla\varphi_i(y) = 2\Big(\mathrm{diag}\left(D^TD\right) + \mathrm{diag}\Big(D^{i^T}D^i\Big) + \mathrm{diag}\Big(\sum\nolimits_{j\neq i} D^{j^T}D^j\Big) + \lambda_2\mathrm{diag}\Big(\sum\nolimits_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\Big)\Big)(x-y) \tag{29}$$

Hence, we obtain:

$$\begin{aligned}\|\nabla\varphi_i(x) - \nabla\varphi_i(y)\| &= \Big\|2\Big(\mathrm{diag}\left(D^TD\right) + \mathrm{diag}\Big(D^{i^T}D^i\Big) + \mathrm{diag}\Big(\sum\nolimits_{j\neq i} D^{j^T}D^j\Big) + \lambda_2\mathrm{diag}\Big(\sum\nolimits_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\Big)\Big)(x-y)\Big\| \\ &\le 2\Big(\mathrm{diag}\left(D^TD\right) + \mathrm{diag}\Big(D^{i^T}D^i\Big) + diag\Big(\sum\nolimits_{j\neq i} D^{j^T}D^j\Big) + \lambda_2\mathrm{diag}\Big(\sum\nolimits_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\Big)\Big)\|\|(x-y)\| \\ &\le 2\left\|\lambda_{max}^1 + \lambda_{max}^2 + \lambda_{max}^3 + \lambda_2\lambda_{max}^4\right\|\|(x-y)\|\end{aligned} \tag{30}$$

where $\lambda_{max}^1 = \lambda_{\max}(\mathrm{diag}(D^TD))$, $\lambda_{max}^2 = \lambda_{\max}(\mathrm{diag}(D^{iT}D^i))$, $\lambda_{max}^3 = \lambda_{\max}(diag(\sum_{j\neq i} D^{jT}D^j))$ and $\lambda_{\max}^4 = \lambda_{\max}\left(\mathrm{diag}\left(\sum_{j\neq i}\tilde{\chi}_j\tilde{\chi}_j^T\right)\right)$. So the (smallest) Lipschitz constant of the gradient $\nabla\varphi_i(\chi_i)$ is $L(\varphi_i) = 2(\lambda_{max}^1 + \lambda_{max}^2 + \lambda_{max}^3 + \lambda_2\lambda_{max}^4)$.

Therefore, we claim that $\boldsymbol{\varphi_i(Z_i)}$ is continuously differentiable with Lipschitz continuous gradient $\boldsymbol{L(\varphi_i)}$.

References

1. Aharon M, Elad M, Bruckstein A (2006) K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(1):4311–4322
2. Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imag Sci 2(1):183–202
3. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. Pattern Anal Mach Intell IEEE Trans 19(7):711–720
4. Bengio S, Pereira F, Singer Y, Strelow D (2009) Group sparse coding. In: Proceedings of the Neural Information Processing Systems
5. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, New York
6. Bryt O, Elad M (2008) Compression of facial images using the k-svd algorithm. J Vis Commun Image Represent 19(4):270–282
7. Cai S, Zuo W, Zhang L, Feng X, Wang P (2014) Support vector guided dictionary learning. In: Computer Vision–ECCV. pp 624–639
8. Candès EJ et al (2006) Compressive sampling. In: Proceedings of the international congress of mathematicians, vol. 3. Madrid, Spain, pp 1433–1452
9. Castrodad A, Sapiro G (2012) Sparse modeling of human actions from motion imagery. Int J Comput Vis 100:1–15
10. Elad M, Aharon M Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans Image Process 15(12):3736–3745

11. Elad M, Aharon M (2006) Image denoising via learned dictionaries and sparse representation. In: Computer Vision and Pattern Recognition, vol. 1. pp 895–900
12. Feng Z, Yang M, Zhang L, Liu Y, Zhang D (2013) Joint discriminative dimensionality reduction and dictionary learning for face recognition. Pattern Recogn 46(8):2134–2143
13. Georghiades A, Belhumeur P, Kriegman D (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660
14. Guha T, Ward RK (2012) Learning sparse representations for human action recognition. IEEE Trans Pattern Anal Mach Learn 34(8):1576–1888
15. Hoyer PO (2002) Non-negative sparse coding. In: Proceedings of the IEEE Workshop Neural Networks for Signal Processing
16. Huang K, Aviyente S (2006) Sparse representation for signal classification. In: Advances in neural information processing system. pp 609–616
17. Jenatton R, Mairal J, Obozinski G, Bach F (2011) Proximal methods for hierarchical sparse coding. J Mach Learn Res 12:2234–2297
18. Jiang ZL, Zhang GX, Davis LS (2012) Submodular dictionary learning for sparse coding. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition
19. Jiang ZL, Lin Z, Davis LS (2013) Label consistent K-SVD: learning a discriminative dictionary for recognition. IEEE Trans Pattern Anal Mach Intell 34:533
20. Kong S, Wang DH (2012) A dictionary learning approach for classification: Separating the particularity and the commonality. In: Proceedings of the European Conference on Computer Vision
21. Mairal J, Elad M, Sapiro G (2008a) Sparse representation for color image restoration. Image Process IEEE Trans 17(1):53–69
22. Mairal J, Bach F, Ponce J, Sapiro G, Zissserman A (2008b) Learning discriminative dictionaries for local image analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
23. Mairal J, Leordeanu M, Bach F, Hebert M, Ponce J (2008c) Discriminative sparse image models for class-specific edge detection and image interpretation. In: Proceedings of the European Conference on Computer Vision
24. Mairal J, Bach F, Ponce J, Sapiro G, Zisserman A (2009) Supervised dictionary learning. In: Proceedings of the Neural Information and Processing Systems
25. Mairal J, Bach F, Ponce J (2012) Task-driven dictionary learning. IEEE Trans Pattern Anal Mach Intell 34(4):791–804
26. Martinez A, Benavente R (1998) The AR face database, CVC Technical Report 24
27. Niyogi X (2004) Locality preserving projections. In: Neural information processing systems, vol. 16. MIT, p 153
28. Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: a strategy employed by v1? Vis Res 37(23):3311–3325
29. Olshausen BA et al (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583):607–609
30. Petrou M, Bosdogianni P (1999) Image processing: the fundamentals. Wiley
31. Pham D, Venkatesh S (2008) Joint learning and dictionary construction for pattern recognition. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition
32. Qiu Q, Jiang ZL, Chellappa R (2011) Sparse dictionary-based representation and recognition of action attributes. In: Proceedings of the International Conference on Computer Vision
33. Ramirez I, Sprechmann P, Sapiro G (2010) Classification and clustering via dictionary learning with structured incoherence and shared features. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. IEEE, 2010, pp 3501–3508
34. Rodriguez F, Sapiro G (2007) Sparse representation for image classification: Learning discriminative and reconstructive nonparametric dictionaries. Preprint: IMA, p 2213
35. Rodriguez M, Ahmed J, Shah M (2008) A spatio-temporal maximum average correlation height filter for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
36. Sadanand S, Corso JJ (2012) Action bank: a high-level representation of activeity in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
37. Sprechmann P, Sapiro G (2010) Dictionary learning and sparse coding for unsupervised clustering. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing
38. Szabo Z, Poczos B, Lorincz A (2011) Online group-structured dictionary learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
39. Turk M, Pentland AP et al (1991) Face recognition using eigenfaces. In: Computer Vision and Pattern Recognition. pp. 586–591
40. Wagner A, Wright J, Ganesh A, Zhou Z, Mobahi H, Ma Y (2012) Toward a practical face recognition system: robust alignment and illumination by sparse representation. Pattern Anal Mach Intell IEEE Trans 34(2):372–386
41. Wang HR, Yuan CF, Hu WM, Sun CY (2012) Supervised class-specific dictionary learning for sparse modeling in action recognition. Pattern Recogn 45(11):3902–3911

42. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009a) Robust face recognition via sparse representation. Pattern Anal Mach Intell IEEE Trans 31(2):210–227
43. Wright JS, Nowak DR, Figueiredo TAM (2009b) Sparse reconstruction by separable approximation. IEEE Trans Signal Process 57(7):2479–2493
44. Wu YN, Si ZZ, Gong HF, Zhu SC (2010) Learning active basis model for object detection and recognition. Int J Comput Vis 90:198–235
45. Yang M, Zhang L (2010) Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In: Computer Vision–ECCV 2010. Springer, pp 448–461
46. Yang JC, Yu K, Huang T (2010a) Supervised translation-invariant sparse coding. In: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition
47. Yang M, Zhang L, Yang J, Zhang D (2010b) Metaface learning for sparse representation based face recognition. In: Proceedings of the IEEE Conference on Image Processing
48. Yang M, Zhang L, Yang J, Zhang D (2011a) Robust sparse coding for face recognition. In: Computer Vision and Pattern Recognition (CVPR). pp 625–632
49. Yang M, Zhang L, Feng XC, Zhang D (2011b) Fisher discrimination dictionary learning for sparse representatio. In: Proceedings of the International Conference on Computer Vision
50. Yang M, Zhang L, Feng XC, Zhang D (2014) Sparse representation based fisher discrimination dictionary learning for image classification. Int J Comput Vis 109(3):209–232
51. Yao A, Gall J, Gool LV (2010) A hough transform-based voting framework for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
52. Zhang Q, Li B (2010) Discriminative k-svd for dictionary learning in face recognition. In: Computer Vision and Pattern Recognition (CVPR). pp 2691–2698
53. Zhang L, Yang M, Feng Z, Zhang D (2010) On the dimensionality reduction for sparse representation based face recognition. In: Pattern Recognition (ICPR), 2010 20th International Conference on IEEE. pp 1237–1240
54. Zhou N, Fan JP (2012) Learning inter-related visual dictionary for object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
55. Zhou MY, Chen HJ, Paisley J, Ren L, Li LB, Xing ZM et al (2012) Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images. IEEE Trans Image Process 21(1):130–144

Bao-Qing Yang received his B.S., M.S., degrees in the School of Internet of Things Engineering, Jiangnan University, in 2005 and 2008, respectively. He is currently a Ph.D. candidate at Department of Automation, Shanghai Jiao Tong University, China. His major research interests include visual surveillance, pattern analysis, sparse representation and face recognition.

Chao-Chen Gu is currently a Research Assistant Professor in School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. He received his bachelor degree from Shandong University, Jinan, China, in 2007, and the Ph.D. degree in Mechanical Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2013. His current research interests include industry robotics, machine vision, and man–machine interfaces.

Kai-Jie Wu is an Associated Professor at School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. He received his Ph.D. degree in Biomedical Engineering from Tianjin University, Tianjin, China, in 2006. His current research explores biomedical optical imaging, medical information processing, and pattern recognition.

Tao Zhang received his Bachelor degree from Henan Polytechnic University, China, in 2008. He is currently a Ph.D. candidate at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. His major research interests include visual surveillance, object detection, and pattern analysis.

Xin-Ping Guan received the B.S. degree in mathematics from Harbin Normal University, Harbin, China, and the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering, both from Harbin Institute of Technology, in 1986, 1991, and 1999, respectively. He is currently a Professor of Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China. He is the (co)author of more than 200 papers in mathematical, technical journals, and conferences. He is the Special appointment professor of Cheung Kong Scholars Programme. His current research interests include functional differential and difference equations, robust control and intelligent control for time-delay systems, chaos control and synchronization, and congestion control of networks.