Multimed Tools Appl (2018) 77:17023–17041
https://doi.org/10.1007/s11042-017-5269-6

Synthesis K-SVD based analysis dictionary learning for pattern classification

Qianyu Wang[1] · Yanqing Guo[1] · Jun Guo[2] · Xiangwei Kong[1]

Received: 19 April 2017 / Revised: 18 September 2017 / Accepted: 28 September 2017 /
Published online: 17 October 2017

Abstract In the fields of computer vision and pattern recognition, dictionary learning techniques have been widely applied. In classification tasks, synthesis dictionary learning is usually time-consuming during the classification stage because of the sparse reconstruction procedure. Analysis dictionary learning, which is another research line, is more favorable due to its flexible representative ability and low classification complexity. In this paper, we propose a novel discriminative analysis dictionary learning method to enhance classification performance. Particularly, we incorporate a linear classifier and the supervised information into the traditional analysis dictionary learning framework by adding a discrimination error term. A synthesis K-SVD based algorithm which can effectively constrain the sparsity is presented to solve the proposed model. Extensive comparison experiments on benchmark databases validate the satisfactory performance of our method.

Keywords Image classification · Dictionary learning · Analysis dictionary learning · Synthesis K-SVD

✉ Yanqing Guo
guoyq@dlut.edu.cn

Qianyu Wang
qianyuw@mail.dlut.edu.cn

Jun Guo
eeguojun@outlook.com

Xiangwei Kong
kongxw@dlut.edu.cn

[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China

[2] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China

1 Introduction

Pattern recognition techniques are indispensable in many real-world applications [10, 11, 14]. Because of the increasing application demands in various areas, methods for handling classification problems are booming. Although pattern classification in different areas have their own technical specialties, the core problem is uniform, which is to construct a favorable model by the known data to handle the coming unknown signals. The flowchart of solving a supervised classification problem is given in Fig. 1. The sample features and sample labels are used to train a model. Unknown samples are classified by the learnt model. Therefore, how to obtain a discriminative model to better classify samples with unknown category information is the main problem behind the pattern classification tasks.

Sparse representation [28] is a very popular research area in the last decade. The aim of this research theme is to represent an instance concisely by a few atoms in a given dictionary for achieving information contained in the data. Sparse prior has deep theoretical bases [1, 19, 20] and extensive practical applications [9, 12, 13, 18], which is the key infrastructure of the sparse representation theory. The dictionaries in early sparse representation methods are usually predefined by mathematical models [24] or all the training data [28], which explores the characteristics existing in the data inadequately. Therefore, dictionary learning (DL) methods which possess learning procedures (i.e., optimizing the dictionary) are proposed.

Sparse prior based DL is a widely used framework in solving diversified computer visual problems [8, 15, 29, 30]. The core idea behind the learning methods is that the structure information of complex original images is efficiently extracted from data itself rather than defined with mathematical models. Actually, a learnt target dictionary often transcends predefined bases in pattern classification tasks, benefitting from better flexibility and adaptability to specific data. Therefore, under the assumption of sparsity, methods for adaptively learning a dictionary with specific properties from instances of data is still an intensively ongoing research theme in DL.

Synthesis dictionary learning (SDL) is a popular strategy in DL research. The synthesis dictionary is able to be obtained by solving an approximate reconstruction problem, which represents images with the linear combination of a few dictionary atoms. Analysis dictionary learning (ADL), which is viewed as a dual viewpoint of SDL, has a more intuitive meaning of the dictionary. The analysis dictionary effects more like a feature transformation matrix which maps image samples into a coding coefficient space. In ADL, a dictionary learned with task specific priors impels the coefficients to possess task specific characteristics in the coding space.

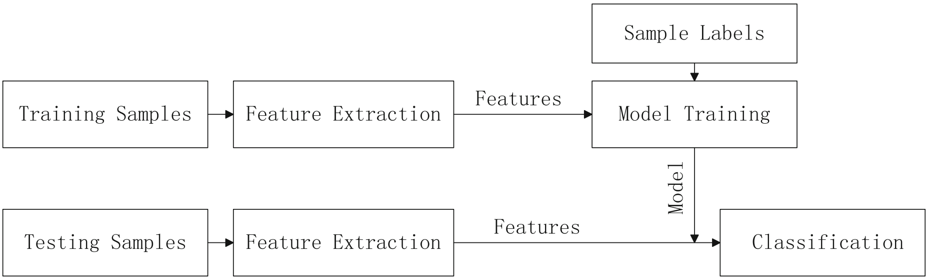


Fig. 1 The flowchart of solving the pattern classification problem

ADL aims to generate coefficients by efficient linear projection, and sparsity constraint makes coefficients concise but informative. Recent years have witnessed that some ADL based methods have been developed. Nam et al. [17] present theoretical insights into analysis models applied in compressed sensing. An analysis K-SVD (singular value decomposition) algorithm [23] is proposed, which is a dual framework of the seminal synthesis K-SVD [1]. Well conditioned transformations proposed by Ravishankar and Bresler [20] have better performance than analytical transforms in image de-noising applications. Afterwards, they enhance this method by adding full-rank and incoherence constraints on the analysis dictionary [19]. To ensure the sparsity of the coding coefficients, Rubinstein and Elad [22] constrain the analysis codes with an imposed hard thresholding operator. However, the works mentioned above do not focus on pattern classification problems.

Concerning learning dictionaries in the tasks of pattern classification, the synthesis framework based DL methods such as D-KSVD [31] and LC-KSVD [6] have reached high accuracies. But the reconstruction procedure is essential in the classification schemes of synthesis methods, leading to classifying samples time-consuming. Moreover, the dictionary that is constructed in this way is optimal neither for reconstructive tasks nor for the discriminative tasks [2]. Though there is a favorable trend of ADL research in recent years, there still exist few researches solving the pattern classification theme by ADL framework. Previous ADL works emphasize presentative ability more while paying no attention on the potential of discrimination power. Because of the flexible representation ability of the transformation matrix, ADL has a necessary discriminability advantage than SDL when dealing with classification tasks. Therefore, we take an exploration for the capability of the analysis framework based method in handling pattern classification problems.

For better classification accuracy and speed, this paper presents a novel pattern classification method, which includes an ADL framework based learning model and an iterative algorithm solving the proposed model. In the model, we incorporate the supervised information into the basic analysis framework and jointly learn a classifier with the dictionary for more discriminability.

Main contributions of this paper are as follows:

1. We propose a novel pattern classification method based on ADL framework, which can avoid the time-consuming reconstruction in classification stage and enhance classification speed. Furthermore, semantic information is incorporated into the model by adding a classification error term.
2. We develop an algorithm solving the objective function of our proposed method based on the synthesis K-SVD. The coefficients and the linear classifier are updated simultaneously in the optimization procedure.
3. Promising results from extensive comparison experiments on five benchmark databases validate the effectiveness of our method and demonstrate the great potential of ADL in pattern classification tasks.

The rest of the paper is organized as follows. In Section 2, the basic frameworks of SDL and ADL are reviewed. Related works are also reviewed in this section. Section 3 presents the proposed ADL method and Section 4 shows the synthesis K-SVD based optimization algorithm for solving the proposed model. We analyze the algorithm complexity after that. Classification scheme is described in Section 5 and comparative studies are reported in Section 6. Finally, the conclusion and our future work are summarized in Section 7.

2 Background

In this section, we will briefly introduce the basic SDL and ADL frameworks. Several works which are related to our method are also described, including sparse representation based classifier (SRC) [28], collaborative representation based classifier (CRC) [32], discriminative K-SVD (D-KSVD) [31], label consistent K-SVD (LC-KSVD) [6] and support vector machine [3] based basic ADL (ADL+SVM) [25].

2.1 Synthesis dictionary learning

The main idea of SDL is to approximately reconstruct the original samples by the combination of dictionary atoms with respective weight factors. The weight factors are stored in the form of coefficients. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] \in \mathbb{R}^{m\times n}$ be the original data matrix, each column of which represents the m-dimensional feature of one sample. Here n denotes the number of samples. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{c\times n}$ be the coding coefficients of $\mathbf{Y}$ over the dictionary and let c be the dimension of the coefficients. The basic problem of SDL is presented as

$$\begin{aligned} \min_{\mathbf{D},\mathbf{X}} \;& \|\mathbf{Y} - \mathbf{DX}\|_F^2 \\ s.t. \;& \mathbf{D} \in \mathcal{A}, \\ & \|\mathbf{x}_i\|_0 \le T_0,\ \forall i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{m\times c}$ is the synthesis dictionary. $\mathcal{A}$ denotes a set of constraints on $\mathbf{D}$ to make the solution non-trivial. And T_0 is the sparsity level.

Euclidean distance is a commonly used tool for the distance metric due to its simplicity [27]. In the above problem, minimizing the square of Frobenius norm means minimizing the sum of all the distances between original data and their approximate representation, i.e., minimizing the residual.

In recent years, the SDL framework based methods have received much interest. SRC [28] minimizes l_1 norm to constrain the sparsity of the coefficients. The l_1 norm is the closest convex approximation of the l_0 norm constraint. The objective function is

$$\begin{aligned} \min_{\mathbf{x}_i} \;& \|\mathbf{x}_i\|_1 \\ s.t. \;& \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 \le \varepsilon \end{aligned} \tag{2}$$

or $\min_{\mathbf{x}_i} \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda\|\mathbf{x}_i\|_1$, where ε is a small constant and λ is a parameter controlling the weight of the corresponding term. This method uses all the training data to construct the dictionary and then classifies testing images by the coefficients. When the training data is huge, this method will result in enormous time consumption in the testing stage.

CRC [32] declares that collaborative representation is also powerful when classifying images. The idea behind it is that different classes share similarities. Some samples from other classes may help to represent a testing sample when lacking of training samples. The CRC is formulated as

$$\min_{\mathbf{x}_i} \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda\|\mathbf{x}_i\|_2^2. \tag{3}$$

Same as SRC, the dictionary of CRC also comprises of all the data. So the same problem exists in CRC. When testing images, the reconstruction error (or representation residual) will be used to make decision.

D-KSVD [31] incorporates the classification error into the basic SDL. The objective function is given as

$$\begin{aligned} \min_{\mathbf{D},\mathbf{W},\mathbf{x}_i} \quad & \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_F^2 + \lambda\|\mathbf{h}_i - \mathbf{W}\mathbf{x}_i\|_F^2 \\ s.t. \quad & \|\mathbf{x}_i\|_0 \leq T_0, \end{aligned} \tag{4}$$

where $\mathbf{h}_i \in \mathbb{R}^l$ is the label of $\mathbf{x}_i$ and l is the number of categories. $\mathbf{W} \in \mathbb{R}^{l\times c}$ is a classifier. In the optimization procedures, D-KSVD uses synthesis K-SVD algorithm to find the optimal solution for the variables.

LC-KSVD [6] adds a label consistency term into the objective function of D-KSVD. The formula of LC-KSVD is defined as

$$\begin{aligned} \min_{\mathbf{D},\mathbf{W},\mathbf{A},\mathbf{x}_i} \quad & \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_F^2 + \lambda\|\mathbf{h}_i - \mathbf{W}\mathbf{x}_i\|_F^2 + \gamma\|\mathbf{g}_i - \mathbf{P}\mathbf{x}_i\|_F^2 \\ s.t. \quad & \|\mathbf{x}_i\|_0 \leq T_0, \end{aligned} \tag{5}$$

where $\|\mathbf{g}_i - \mathbf{P}\mathbf{x}_i\|_F^2$ is the label consistency term. $\mathbf{g}_i \in \mathbb{R}^g$ is the 'ideal' sparse code for the input sample $\mathbf{x}_i$ and $\mathbf{P} \in \mathbb{R}^{g\times c}$ is a linear transformation matrix. LC-KSVD also uses synthesis K-SVD algorithm to optimize the variables.

2.2 Analysis dictionary learning

ADL aims to obtain a projective matrix (i.e., analysis dictionary) to implement the approximately sparse representation in transformed domain. The analysis dictionary $\mathbf{\Omega} \in \mathbb{R}^{c\times m}$, which has an intuitive explanation like feature transformation, is defined as

$$\begin{aligned} \min_{\mathbf{\Omega},\mathbf{X}} \quad & \|\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\|_F^2 \\ s.t. \quad & \mathbf{\Omega} \in \mathcal{D}, \\ & \|\mathbf{x}_i\|_0 \leq T_0, \forall i = 1, \ldots, n, \end{aligned} \tag{6}$$

where $\mathcal{D}$ denotes a set of constraints on $\mathbf{\Omega}$. It can either be matrix with relatively small Frobenius norm or with normalized rows. Sparsity is constrained by parameter T_0. Problem (6) is the basic framework of ADL (referred to as basic ADL later). In dictionary learning, sparsity is advocated as a key prior not only for many synthesis methods, but also for analysis approaches. It can lead to simple and succinct descriptions of natural phenomena.

Seldom works focus on solving the pattern classification theme by ADL. ADL+SVM [25] learns an analysis dictionary by minimizing the representative residual. The basic ADL is based on the criterion function (6). During the training stage of the model, thresolding operator is applied to keep the sparsity of the coefficients. The learnt dictionary is used to obtain coefficients corresponding to all the samples. Then, coefficients of training samples are put into the SVM [3] to train a classifier for testing samples.

3 The proposed model

The main idea of problem (6) is to learn an optimized dictionary with which we can gain effective coefficients of every sample. However, this model only focuses on representative power without considering the discrimination ability of the dictionary. For performing pattern classification tasks better, we expect that images from the same class have similar representations. This problem can be turned to learn a dictionary which impels the coefficients from intra-class images similar and close but maximizes the dissimilarity of coefficients from inter-class images.

3.1 Classification error function

To ensure coefficients to have the above properties, we attempt to leverage the supervised information of input samples. But supervised information is hard to be incorporated into the basic ADL model directly. Coefficients obtained from minimizing problem (6) are informative and can be treated as feature vectors in the sparsity coding space for images. In addition, handling the coefficients in transformed domain is more general. Therefore, we introduce a classifier to constrain coefficients with class labels of samples. Let $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n] \in \{0, 1\}^{l \times n}$ be the class label matrix. Note that the nonzero place of $\mathbf{h}_i = [0, 0, \ldots, 1, \ldots, 0, 0]^T$, which is one column of $\mathbf{H}$, indicates the class of $\mathbf{y}_i$. A well performed classifier is often obtained by optimizing some functions, which have a generalized form as

$$\min_{\mathbf{W}} \sum_i f\{\mathbf{h}_i, \mathbf{x}_i, \mathbf{W}\} + \delta \|\mathbf{W}\|_F^2, \tag{7}$$

where f is the classification loss function. $\|\mathbf{W}\|_F^2$ is a regularization term making the solution of the classifier $\mathbf{W} \in \mathbb{R}^{l \times c}$ non-trival and δ is a parameter controlling its weight. Here, we set the classifier to be linear and define a transformation $g(\mathbf{W}, \mathbf{x}_i) = \mathbf{W}\mathbf{x}_i$. The deviation of $\mathbf{W}\mathbf{x}_i$ from its target form $\mathbf{h}_i$ can be measured with l_2 norm. Based on these, the linear classifier can be obtained by solving the following classification error function:

$$\min_{\mathbf{W}} \sum_i \|\mathbf{h}_i - \mathbf{W}\mathbf{x}_i\|_2^2 + \delta \|\mathbf{W}\|_F^2. \tag{8}$$

3.2 Adding discriminability to ADL

The learnt dictionary may be sub-optimal for classification if we learn the dictionary and the classifier individually. It is better to optimize problem (6) and (8) jointly. Therefore, we combine the classification error function with the basic ADL framework to achieve optimal solutions in classification. The proposed model can be formulated as

$$\begin{aligned} \min_{\boldsymbol{\Omega}, \mathbf{X}, \mathbf{W}} \quad & \|\mathbf{X} - \boldsymbol{\Omega}\mathbf{Y}\|_F^2 + \alpha \|\mathbf{H} - \mathbf{W}\mathbf{X}\|_F^2 \\ s.t. \quad & \boldsymbol{\Omega} \in \mathcal{D}, \\ & \|\mathbf{x}_i\|_0 \le T_0, \ \forall i = 1, \ldots, n, \end{aligned} \tag{9}$$

where α is a positive scalar controlling the contribution of the classification error term. $\|\mathbf{X} - \boldsymbol{\Omega}\mathbf{Y}\|_F^2$ is called sparsification error or representation error. It shows the disparity between image representations in the transformed space and the coefficients with target sparsity level. The classification error term $\|\mathbf{H} - \mathbf{W}\mathbf{X}\|_F^2$ represents the integral degree of approximation between the representation of transformed $\mathbf{X}$ in transformation domain and the desired coding form $\mathbf{H}$. In detail, this term enforces data from each class to have their own different target expressions after transformation. That is to say, the coefficients of one sample are similar with ones from identical category and distinct with ones from other classes. An appropriate value of parameter α can make our model balance with representative power and discriminative power.

Integrating the classification error term with the representation error means the supervised information is incorporated in the learning period. This integration enforces the learnt dictionary to have more capability of discrimination and to generate discriminative coefficients when classifying samples with unknown class information. In the following section, we will describe a strategy to solve the optimization problem (9).

The overview of the idea of our proposed method is presented in Fig. 2. As the figure shows, the analysis dictionary transforms the features of samples into a sparse space. The coefficients in the sparse space hold the property that only a portion of elements in each column of the coefficient matrix is nonzero. After that, a classifier makes the representations of the samples lie in the label space (i.e., predictive vectors), making it easy to classify samples by finding the closest coordinate axis of the vector.

Compared to SRC and CRC which just use features of all the samples as the dictionary, our proposed method has a training stage which optimizes the dictionary so that we can get comparable classification accuracy even under a smaller dictionary (which will be shown in Section 6). Although D-KSVD and LC-KSVD have dictionary optimization procedure, the dictionary reconstructed in the synthesis way is optimal neither for reconstructive tasks nor for the discriminative tasks. Our proposed method is based on the ADL framework so that its performance does not rely on the dictionary size as the SDL based method. The ADL+SVM is an ADL based method designed for classification tasks. However, the discrimination ability is not fully exploited for separately learning dictionary and classifier in the training stage. Our proposed method incorporates the semantic labels into the objective function so that the dictionary we learnt is more discriminative.

4 Algorithm

4.1 Optimization procedures

Because of the constraint of l_0 norm, the overall objective function in (9) is non-convex. One existing optimization method for ADL is to compute analytical solution to each variable with the sparsity implemented by a soft or hard sparsity operator. However, the sparsity operator is too coarse. Setting relatively small values of coefficients to be zero roughly might omit some important information in the training stage. The seminal synthesis K-SVD applies SVD to learn the dictionary and coefficients, with effectively constraining

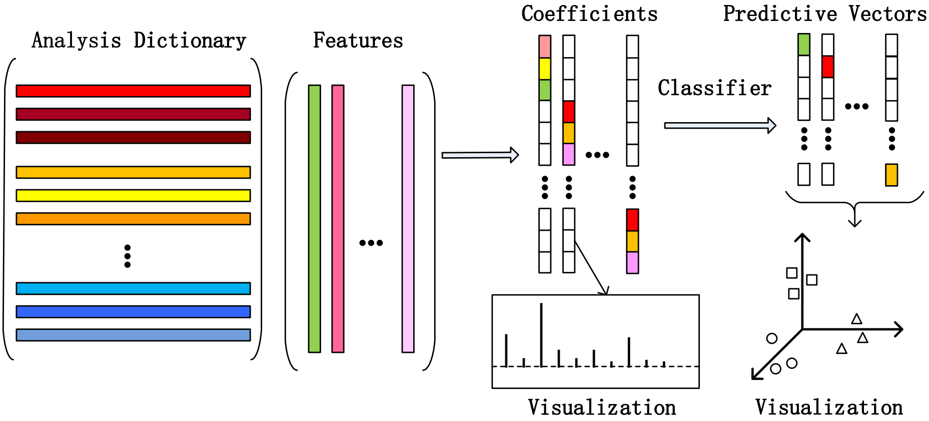


Fig. 2 The overview of our proposed ADL method

the sparsity. But previous ADL based works related to the synthesis K-SVD employ pseudo inverse of analysis dictionary to satisfy the entrance requirements of the K-SVD. However, the equivalency of the variant form of pseudo inverse and the model prototype is intangible in this case.

Illuminated by some works in SDL, we introduce a new strategy for solving our proposed model. Though based on the synthesis K-SVD, it has a large difference with the pseudo inverse method. Gradient technique is also incorporated into the optimization procedure. The iterative optimization algorithm contains the following steps:

Update {X, W} Fixing the dictionary $\mathbf{\Omega}$, the solutions for $\mathbf{X}$ and $\mathbf{W}$ in the present iteration can be obtained concurrently by utilizing the synthesis K-SVD. Problem (9) can be rewritten in the form of amalgamation of all the terms, which is

$$\begin{aligned} \min_{\mathbf{X},\mathbf{W}} \quad & \left\| \begin{pmatrix} \mathbf{\Omega Y} \\ \sqrt{\alpha}\mathbf{H} \end{pmatrix} - \begin{pmatrix} \mathbf{R} \\ \sqrt{\alpha}\mathbf{W} \end{pmatrix} \mathbf{X} \right\|_F^2 \\ s.t. \quad & \|\mathbf{x}_i\|_0 \le T_0, \forall i = 1, \ldots, n, \end{aligned} \tag{10}$$

where $\mathbf{R}$ is initialized as an identity matrix. Denote the two combined matrices in the parentheses of problem (10) as

$$\mathbf{Y}_{new} = (\mathbf{Y}^T\mathbf{\Omega}^T, \sqrt{\alpha}\mathbf{H}^T)^T, \tag{11}$$

$$\mathbf{Q}_{new} = (\mathbf{R}^T, \sqrt{\alpha}\mathbf{W}^T)^T. \tag{12}$$

In the original K-SVD algorithm, the matrix $\mathbf{Q}_{new}$ is column-wise l_2 normalized. The optimization problem (10) is equivalent to the following problem:

$$\begin{aligned} < \mathbf{Q}_{new}, \mathbf{X} > \; = \; & \arg \min_{\mathbf{Q}_{new},\mathbf{X}} \|\mathbf{Y}_{new} - \mathbf{Q}_{new}\mathbf{X}\|_F^2 \\ & s.t. \;\; \|\mathbf{x}_i\|_0 \le T_0, \; \forall i = 1, \ldots, n. \end{aligned} \tag{13}$$

The optimization problem in (13) can be solved by K-SVD, with entrance parameters being $\mathbf{Y}_{new}$, $\mathbf{Q}_{new}$ and T_0, etc. Particularly, let $\mathbf{x}_R^k$, which is the k-th row of $\mathbf{Y}_{new}$, be the corresponding coefficients of the k-th column of $\mathbf{Q}_{new}$, and denoted as $\mathbf{q}_k$. Let $\mathbf{E}_k = (\mathbf{Y}_{new} - \sum_{j \neq k} \mathbf{q}_j \mathbf{x}_R^j)$, then discard the zero entries in $\mathbf{x}_R^k$ and $\mathbf{E}_k$, with corresponding results marked as $\widetilde{\mathbf{x}}_R^k$ and $\widetilde{\mathbf{E}}_k$. We optimize problem (14) to obtain $\mathbf{q}_k$ and $\widetilde{\mathbf{x}}_R^k$.

$$< \mathbf{q}_k, \widetilde{\mathbf{x}}_R^k >= \arg \min_{\mathbf{q}_k, \widetilde{\mathbf{x}}_R^k} \left\| \widetilde{\mathbf{E}}_k - \mathbf{q}_k \widetilde{\mathbf{x}}_R^k \right\|. \tag{14}$$

Decomposing $\widetilde{\mathbf{E}}_k$ by SVD, we have $\widetilde{\mathbf{E}}_k = \mathbf{U\Sigma V}^T$. Let $\mathbf{q}_k = \mathbf{U}(:, 1)$ and $\widetilde{\mathbf{x}}_R^k = \mathbf{\Sigma}(1, 1)\mathbf{V}$ $(:, 1)$. After that, the nonzero values of $\mathbf{x}_R^k$ are replaced by $\widetilde{\mathbf{x}}_R^k$.

After one iteration, $\mathbf{Q}_{new}^{t+1}$ for next iteration's K-SVD process is directly assigned by $\mathbf{Q}_{new}^{t+1} = \mathbf{Q}_{new}^t$, instead of renewing it with identity matrix and $\mathbf{W}^t$ as (12). That is because the update of other variables is independent w.r.t. $\mathbf{W}$ and the coefficients $\mathbf{X}$ can be directly achieved from the output of K-SVD. In addition, as Fig. 3 shows, $\mathbf{R}$ is still very similar to identity matrix after dozens of iterations. Consequently, taking out $\mathbf{W}$ from $\mathbf{Q}_{new}$ at the end of the training period without extra operations is tolerable, such as constraining $\mathbf{R}$ to be strict identity matrix.

Update Ω Fixing $\mathbf{W}$ and $\mathbf{X}$, the solution for $\mathbf{\Omega}$ is computed through matrix derivation, followed by normalization of the rows of $\mathbf{\Omega}$. For simple computation, we constrain the set $\mathcal{D}$

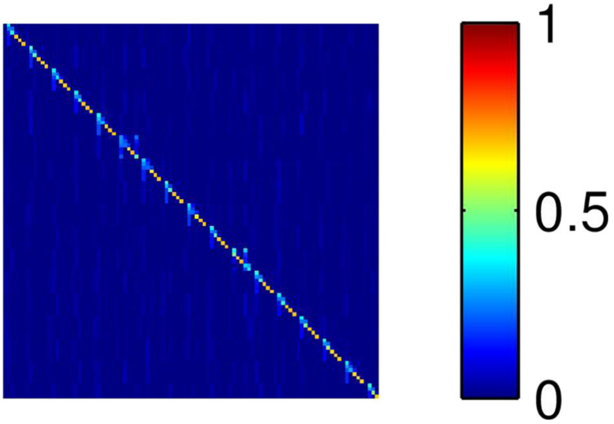


Fig. 3 Values of $\mathbf{R}$ after 30 iterations, obtained from $\mathbf{Q}_{new}^{30}$

to be matrices which have small Frobenius norm. Then we add the following regularization term into the object function. The formulation in this situation is given as

$$\min_{\mathbf{\Omega}} \ \|\mathbf{X}-\mathbf{\Omega}\mathbf{Y}\|_F^2+\alpha\|\mathbf{H}-\mathbf{W}\mathbf{X}\|_F^2+\beta\|\mathbf{\Omega}\|_F^2, \tag{15}$$

where $\beta > 0$ is a parameter which indicates the weight of the penalty term. The role of this term lies in avoiding singularity and over-fitting issues as well as ensuring the stable solution of dictionary. To solve problem (15), we transform the function into the form of trace. After omitting the independent terms, an equivalent problem (16) is obtained.

$$\min_{\mathbf{\Omega}} Tr(\mathbf{Y}^T\mathbf{\Omega}^T\mathbf{\Omega}\mathbf{Y}-2\mathbf{X}^T\mathbf{\Omega}\mathbf{Y}+\beta\mathbf{\Omega}^T\mathbf{\Omega}). \tag{16}$$

Let the first derivative w.r.t. $\mathbf{\Omega}$ be zero and we can obtain the analytical solution for the dictionary as

$$\mathbf{\Omega}=\mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T+\beta\mathbf{I})^{-1}, \tag{17}$$

where $\mathbf{I}$ is an identity matrix. Then we normalize rows of the dictionary, which results in better performance empirically.

4.2 Initialization

Here we illustrate the initialization of our scheme. In ADL, the coefficients and the dictionary have the identical column length and they can mathematically derive each other if one is fixed. Considering the sparsity constraint, the coefficients should have a proper length to avoid accidental error judgments. Let $\mathbf{a} \in \mathbf{1}$ be an all one vector and a be the length of $\mathbf{a}$. And $\mathbf{H}_{ini}$ denotes the initial matrix. Assume $\mathbf{H}_{ini} \in \mathbb{R}^{c\times b}$. To achieve high-dimensional codes, Kronecker product is used for initialization, which is

$$\mathbf{H}_{ini}=\mathbf{H}\otimes\mathbf{a}=\begin{bmatrix}\mathbf{H}_{11}\mathbf{a} & \cdots & \mathbf{H}_{1b}\mathbf{a}\\ \vdots & \ddots & \vdots\\ \mathbf{H}_{a1}\mathbf{a} & \cdots & \mathbf{H}_{ab}\mathbf{a}\end{bmatrix}. \tag{18}$$

The aim of this process is to indirectly obtain a preferable initial dictionary by (17). An analogously procedure can compute an analytical solution w.r.t $\mathbf{W}$ for initialization, which is

$$\mathbf{W} = \alpha \mathbf{H}\mathbf{X}^T(\alpha \mathbf{X}\mathbf{X}^T + \delta \mathbf{I})^{-1}, \tag{19}$$

where the parameter δ is fixed as 10^{-6} for the feasibility of the inverse operation (avoiding the singular solution).

The above procedures are summarized in Algorithm 1.

Algorithm 1 Algorithm for solving our proposed model

```
Input:
    Input Training data $\mathbf{Y}$ and class label $\mathbf{H}$;
    Parameter $a$ about the size of dictionary;
    Regularization parameters $\alpha, \beta$;
    Number of iterations $T$.
Output:
    The analysis dictionary $\mathbf{\Omega}$ and classifier $\mathbf{W}$.
Initialize $\mathbf{X}^{(0)}$ by (18), $\mathbf{\Omega}^{(0)}$ by (17), $\mathbf{W}^{(0)}$ by (19);
for $t = 1$ to $T$ do
    Update the classifier $\mathbf{W}^{(t)}$ and the coefficient matrix $\mathbf{X}^{(t)}$ by solving (10) using synthesis K-SVD [1];
    Update $\mathbf{\Omega}^{(t)}$ by solving (17);
end for
```

4.3 Complexity analysis

The proposed algorithm alternates between the classifier-coefficients updating step and the analysis dictionary updating step. We now discuss their computational costs. Note that $\mathbf{Y} \in \mathbb{R}^{m\times n}$, $\mathbf{X} \in \mathbb{R}^{c\times n}$, $\mathbf{\Omega} \in \mathbb{R}^{c\times m}$, $\mathbf{H} \in \mathbb{R}^{l\times n}$, and $\mathbf{W} \in \mathbb{R}^{l\times c}$. The computational cost of the classifier and coefficients updating step is dominated by the computation of the K-SVD of $\widetilde{\mathbf{E}}_k$. Considering the number of columns of $\mathbf{Q}_{new}$, the cost of this step scales as $O(cn(c + l)^2)$. For the analysis dictionary updating step, the computation of the product $\mathbf{X}\mathbf{Y}^T$ requires $O(T_0cnm)$ multiply operations for an $\mathbf{X}$ with sparsity T_0. The inverse operation of $(\alpha\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}$ costs $O(c^3)$. Therefore, the total cost per iteration of the proposed algorithm scales as $O(cn(c + l)^2)$. The algorithm of LC-KSVD, which learns a label consistent matrix $\mathbf{P} \in \mathbb{R}^{g\times c}$, a classifier $\mathbf{W} \in \mathbb{R}^{l\times c}$ and a synthesis dictionary $\mathbf{D} \in \mathbb{R}^{m\times c}$, scales as $O(cn(m + l + g)^2)$. Compared to LC-KSVD, our proposed method holds a similar computational complexity if under the assumption that $c \approx m \gg l$. But if we set $c \ll m$, the complexity of our algorithm is much lower.

5 Classification

To classify a testing query $\mathbf{y}_i$, its corresponding coefficient $\mathbf{x}_i$ can be seen as a new image representation, which already has a certain extent of discrimination in feature decision level. But it is not obvious enough to directly predict categorical attributes of $\mathbf{x}_i$. Therefore, we acquire a predictive label as the decision level feature for easier decision making. Because our method mainly endeavors to achieve an approximate classification representation of one sample after operations over the analysis dictionary and the classifier, the predictive vector is close to the form of its real label (i.e., the element position indicating the real class has obviously larger value than other nearly-zero positions).

Based on these, preliminary coefficients are obtained via the operation of multiplying $\mathbf{\Omega}$ by the testing data $\mathbf{y}_i$. Keeping in view that the proposed method imposes sparsity on coefficients, a hard thresholding operator is used to maintain the sparse characteristic of $\mathbf{x}_i$. The operator reserves elements with T_0 biggest absolute values and sets the others to be zero. Then let the predictive vector $\mathbf{h}_{i_{pre}} = \mathbf{W}\mathbf{x}_i$. The above process is fast and effective because the learnt dictionary and classifier already have strong discrimination. Compared to LC-KSVD in which the SVD procedure is also necessary in testing stage, our method only needs matrix multiplications and a thresolding process, resulting in much shorter classification time.

The predictive vector $\mathbf{h}_{i_{pre}}$ has a very approximate shape to the corresponding real label vector $\mathbf{h}_i$, with only one element obviously bigger than others. Therefore the location of the largest element in $\mathbf{h}_{i_{pre}}$ is utilized to determine the category. The final problem turns to find out where the element with the maximum value is in the predictive vector.

6 Experiment

We verify the performance of our method on four applications under five databases, including face recognition under Extended Yale face database B (YaleB) [5] and AR face database [16], object recognition under Caltech 101 database [7], scene category under fifteen scene dataset (Scene 15) [26] and action recognition under UCF 50 database [21]. The above databases are widely used in evaluating the performance of sparse representation based classification methods. The features we use are provided by Jiang[1] and Corso.[2] For fair comparison, the experiment settings we follow are in accordance with [6].

6.1 Databases

In Table 1, we list the information about the databases and features we use. The third column shows the types and dimensions of the features. On YaleB and AR face databases, random face features are generated by the projection with a randomly generated matrix. On Caltech 101 and Scene 15 databases, features are achieved by extracting scale invariant feature transform (SIFT) descriptors, max pooling in spatial pyramid and reducing dimensions by principal component analysis (PCA) [26]. On UCF 50 database, action bank features are extracted from the action clips [4]. The forth column contains the numbers of samples of each class in total and in use. Therein Caltech 101 database has 101 object classes and 1 background class. Some typical images from each database are presented in Fig. 4.

6.2 Experiment details

We compare the proposed method with previous famous approaches including SRC [28], CRC [32], K-SVD [1], D-KSVD [31], and LC-KSVD [6]. Meanwhile, to gain more insights

[1] http://www.umiacs.umd.edu/~zhuolin/projectlcksvd.html.

[2] http://www.cse.buffalo.edu/~jcorso/r/actionbank.

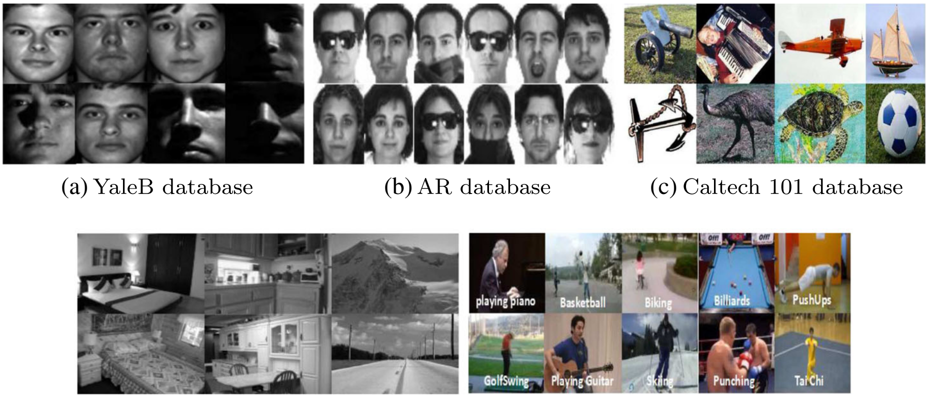
Table 1 Information for five databases and features

	Content	Feature/dimension	Number (total/use)	Class (total/use)	Train/test
YaleB [5]	Face	Random face/504	64/64	38/38	32/32
AR [16]	Face	Random face/540	26/26	126/100	20/6
Caltech101 [7]	Object	SIFT/3,000	31–800/all	102/102	30/rest
Scene15 [26]	Scene	SIFT/3,000	200–400/all	15/15	100/rest
UCF50 [21]	Action	Action bank/5,000	6880/all	50/50	$\frac{4}{5}/\frac{1}{5}$

into how the classification error term affects the testing performance, we implement the baseline experiment ADL+SVM, which solves the basic ADL model and uses SVM based classifier to classify testing images.

The performance of SRC and CRC are measured when the dictionary size is the number of all training samples following the protocol of original works, labeling with (all) in the table. Results of methods other than SRC and CRC are taken under the same size of dictionary with the proposed method.

In our learning strategy, the number of dictionary atoms (i.e., rows of the dictionary) is the integral multiple of the number of classes, which is analogous to the synthesis dictionary. The magnitude of rows of the learnt dictionary is set to be numbers between 500 and 600 according to different database situations. We find that accuracy rates are insensitive to $T_0 \in [40, 70]$ with regard to the dictionary size. Here we fix the sparsity thresholding as 45 for each database. Although a larger sparsity in this range slightly improves the accuracy of class recognition, it is found to result in slowing the optimization process down. Keeping values of sparsity constraint and the number of dictionary size fixed, parameter α and β are tuned by 5-fold cross validation and optimized by using grid search strategy on each database. The grid search is a common strategy to find proper parameters in machine learning, which means that we firstly search in the range of $[10^{-6}, 10^{-5}, \ldots, 10^{5}, 10^{6}]$ and then search a smaller grid with proper interval size determined by preliminary results.



(a) YaleB database (b) AR database (c) Caltech 101 database

(d) Scene 15 database (e) UCF 50 database

Fig. 4 Typical examples from five databases

Table 2 Classification accuracy (%) on different databases

	YaleB	AR	Caltech 101	Scene 15	UCF 50
ADL+SVM	95.4	96.1	64.5	91.8	72.3
K-SVD	93.1	86.5	73.0	86.7	51.5
D-KSVD	94.1	88.8	73.2	89.1	57.8
SRC(all)	96.5	97.5	70.7	91.8	68.4
CRC(all)	97.0	98.0	68.2	92.0	68.6
LC-KSVD	96.7	97.8	73.6	92.9	70.1
Ours	96.9	97.7	74.4	97.4	74.6

6.3 Results and analyzes

The experimental results in terms of the classification accuracy are listed in Table 2. As we can see, on all databases, our method gains notably higher accuracy than ADL+SVM, which proves availability of adding jointly learnt classification error term for discrimination of ADL. Our proposed method has a least improvement of 1.6% over the recognition rate on YaleB database compared with ADL+SVM, and the recognition accuracy improves more obviously on database Caltech 101 and Scene 15.

Our method also achieves favorable results compared with SDL based methods, especially outperforms K-SVD, D-KSVD and LC-KSVD on all the databases about classification accuracy. Furthermore, accuracies of the proposed method for database Caltech 101, Scene 15 and UCF 50 are highest in the listed methods. As for face recognition (YaleB and AR database), results of other methods are already very high, the improvement space is limited. The recognition rates of the CRC in these two databases are a little higher than those of our proposed method. That is mainly because face recognition pays attention to facial details. Collaborate representation holds the details (i.e., coefficients with small values) while sparse representation applied by our method may lose some details due to of the sparsity. In addition, there are fewer variations among faces within the same class on databases YaleB and AR than the other three databases. Our method focuses more on discriminative representation, which may lose some original similarity information. Moreover, dictionaries of SRC and CRC are quite large. The number of dictionary atoms is as same as the number of all training samples. The results in Table 2 also show that learning a much smaller analysis dictionary by our method can gain comparable or better performance than learning large dictionaries in synthesis framework based methods.

In Table 3, we can observe the values of optimized parameters on the five databases. Each database has its own optimized parameter value due to the specific properties of their original features. The parameter α is a weight to balance the representation error term and the classification error term. The larger the α is, the smaller the classification error is. It may

Table 3 Parameters selection in the best performance situation for parameter α and β, determined by grid search strategy

	YaleB	AR	Caltech101	Scene15	UCF50
α	8	6	5	30	30
β	0.03	0.04	5	2	0.02

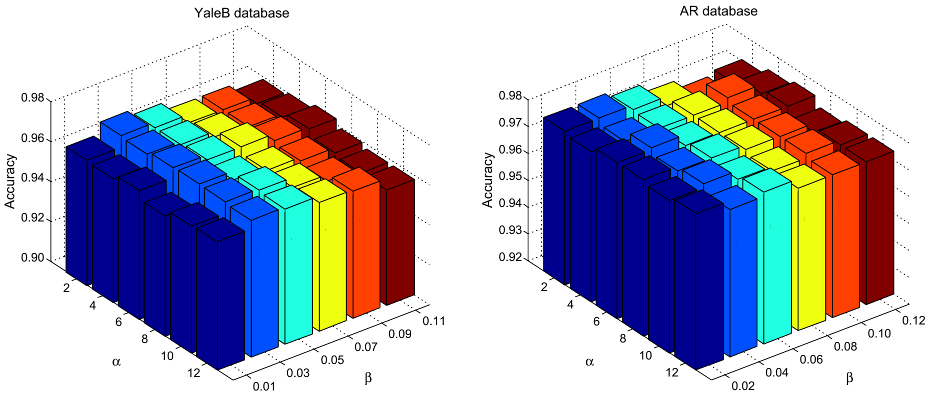


Fig. 5 Results of parameter selection of α and β on recognition rate on YaleB and AR databases

arise the over-fitting problem if it is too big. The optimized values of α are all more than 1. That is because semantic information is significative to get more discriminability. A proper α will make classification error felicitously low so that more discriminative information can be transferred to the dictionary. β is mainly designed to penalize the scale of the dictionary and avoid over-fitting. Therefore, its optimized value changes with properties of raw features. For α and β, their impacts on the classification results are slight in certain range. But if the values of these two parameters are too large or too small, the classification accuracy will decrease. In Fig. 5, 3-D histograms show that the values of accuracy vary as parameter α and β change on YaleB and AR databases. We can observe that good performance is achieved at $\alpha = 8$ and $\beta = 0.03$ on YaleB and $\alpha = 6$ and $\beta = 0.04$ on AR.

Table 4 lists the time cost for testing one image with our method and LC-KSVD on databases YaleB (dictionary size = 540) and AR (dictionary size = 600). Meanwhile, LC-KSVD* learns a model with $\lambda = 0$ in problem (5). Experiments for time are carried out in MATLAB of the same computer with 32 GB memory and 2.6G Hz Intel CPU for fair. The results are calculated through dividing the testing time by the totality of all testing images. Time advantage is fairly obvious. Our method is approximately 7 times faster than LC-KSVD. This advantage mainly owns to the simple classification scheme of ADL which only uses multiplication and a sparsity operator, without the time costing iterative reconstruction processing of SDL. This is a great advantage of ADL compared to SDL. The accuracies and time costs in tables can demonstrate that analysis dictionary learning has huge potential in pattern classification tasks.

Table 4 Time (ms) for classifying one testing image

	YaleB	AR
LC-KSVD*	0.88	0.91
LC-KSVD	0.89	0.94
Ours	0.11	0.13

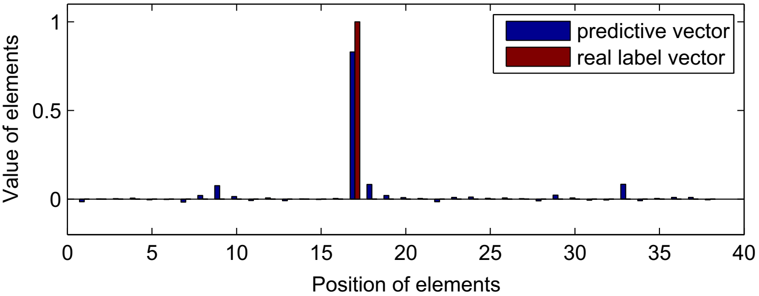
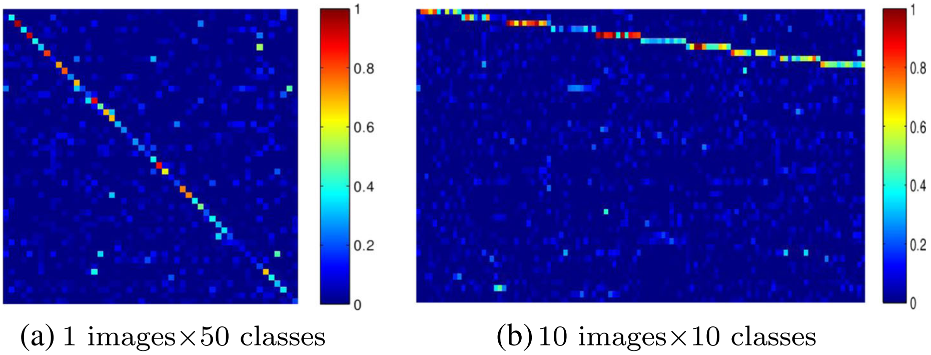


Fig. 6 Output $\mathbf{h}_{i_{pre}}$ (predictive vector, the blue bars) and the real label vector (the red bars) of a testing image from YaleB database

In Fig. 6, the two most closed blue and red bars are referring to elements corresponding to a same position. The blue bars represent the predictive label of a random sample in YaleB database. And the red bars belong to the real label vector of the same sample. As Fig. 6 shows, $\mathbf{h}_{i_{pre}}$ has a very approximate shape to $\mathbf{h}_i$ with only one entry obviously bigger than the others. Thus, the location of the largest element in $\mathbf{h}_{i_{pre}}$ indicates the class of one sample.

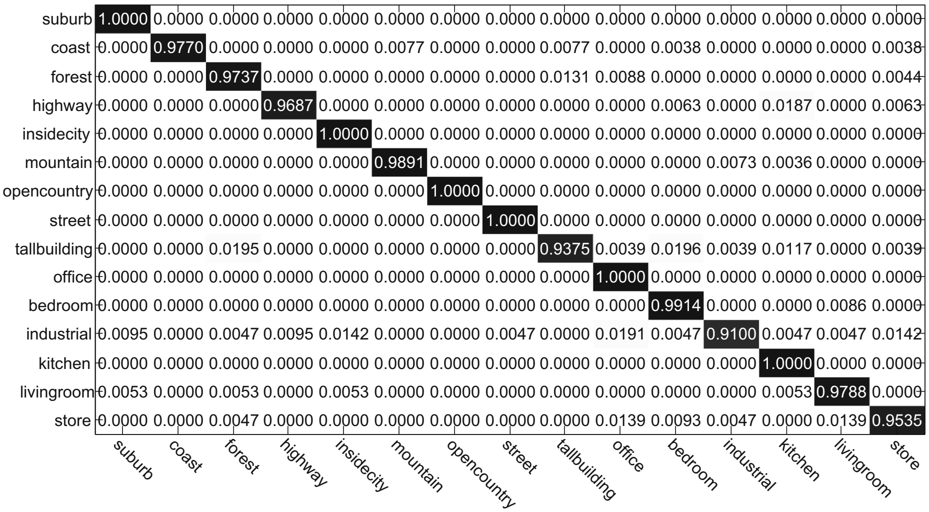
Figure 7 shows the magnitude of every elements in $\mathbf{H}_{i_{pre}}$ of some testing images in UCF 50. The larger the value of matrix element is, the more reddish the color is. It can be seen that the lightest points in the columns decide the category to be classified to and some reddish points deviated from others will lead to the wrong classification.

Figure 8 is the confusion matrix for all the testing samples on Scene 15 database. It presents proportion of images in each category classified to all categories. The sum of each row is 1. We can observe that most images can be classified into the right category, with some class even getting all right classification. From the figures, we can conclude that the desired effect of our proposed method is reached.



(a) 1 images×50 classes (b) 10 images×10 classes

Fig. 7 Final $\mathbf{H}_{i_{pre}}$, calculated on part of testing image randomly selected from UCF 50 database



	suburb	coast	forest	highway	insidecity	mountain	opencountry	street	tallbuilding	office	bedroom	industrial	kitchen	livingroom	store
suburb	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
coast	0.0000	0.9770	0.0000	0.0000	0.0000	0.0077	0.0000	0.0000	0.0077	0.0000	0.0038	0.0000	0.0000	0.0000	0.0038
forest	0.0000	0.0000	0.9737	0.0000	0.0000	0.0000	0.0000	0.0000	0.0131	0.0088	0.0000	0.0000	0.0000	0.0000	0.0044
highway	0.0000	0.0000	0.0000	0.9687	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0063	0.0000	0.0187	0.0000	0.0063
insidecity	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
mountain	0.0000	0.0000	0.0000	0.0000	0.0000	0.9891	0.0000	0.0000	0.0000	0.0000	0.0000	0.0073	0.0036	0.0000	0.0000
opencountry	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
street	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
tallbuilding	0.0000	0.0000	0.0195	0.0000	0.0000	0.0000	0.0000	0.0000	0.9375	0.0039	0.0196	0.0039	0.0117	0.0000	0.0039
office	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bedroom	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9914	0.0000	0.0000	0.0086	0.0000
industrial	0.0095	0.0000	0.0047	0.0095	0.0142	0.0000	0.0000	0.0047	0.0000	0.0191	0.0047	0.9100	0.0047	0.0047	0.0142
kitchen	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
livingroom	0.0053	0.0000	0.0053	0.0000	0.0053	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0053	0.9788	0.0000
store	0.0000	0.0000	0.0047	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0139	0.0093	0.0047	0.0000	0.0139	0.9535

Fig. 8 Confusion matrix on database Scene 15

7 Conclusion

In this paper, we propose a novel pattern classification method which is based on analysis dictionary learning and synthesis K-SVD. In the proposed model, to improve the discriminability of ADL, a classification error term is introduced into the basic ADL framework. To effectively solve the proposed objective function, an iterative algorithm based on synthesis K-SVD is developed. In the optimization procedure, the classifier and the coefficient matrix can be updated simultaneously. We test the performance on five different databases. Experimental results demonstrate that our analysis dictionary learning method can gain comparable or better classification results than previous DL methods. The time of classifying images by our method is much shorter than that by LC-KSVD. The results also show that the synthesis K-SVD can validly solve the ADL optimization problems.

Nowadays, more and more visual data can be described by different views. We believe that dictionary learning is not limited to single-view tasks, and can be used in multi-view tasks. For future work, we plan to explore the ability of analysis dictionary learning in the multi-view classification problems.

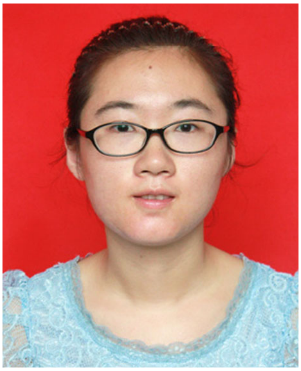
Acknowledgements This work is funded by the National Natural Science Foundation of China (Grant No. 61402079), the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR, No. 201600022).

References

1. Aharon M, Elad M, Bruckstein AK (2006) K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(11):4311–4322
2. Bahrampour S, Nasrabadi N, Ray A, Jenkins W (2016) Multimodal task-driven dictionary learning for image classification. IEEE Trans Image Process 25(1):24
3. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27:1–27:27

4. Corso JJ (2012) Action bank: a high-level representation of activity in video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1234–1241
5. 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
6. Jiang Z, Lin Z, Davis LS (2013) Label consistent K-SVD: learning a discriminative dictionary for recognition. IEEE Trans Pattern Anal Mach Intell 35(11):2651–2664
7. Li F, Rob F, Pietro P (2007) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Comput Vis Image Underst 106(1):59–70
8. Li Y, Guo Y, Guo J, Li M, Kong X (2015) CRF With locality-consistent dictionary learning for semantic segmentation. In: Third IAPR asian conference pattern recognition, ACPR 2015. Kuala Lumpur, Malaysia, pp 509–513
9. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2activity: recognizing complex activities from sensor data. In: Proceedings of the twenty-fourth international joint conference on artificial intelligence, IJCAI. Buenos Aires, Argentina, pp 1617–1623
10. Liu L, Cheng L, Liu Y, Jia Y, Rosenblum DS (2016) Recognizing complex activities by a probabilistic interval-based model. In: Proceedings of the thirtieth AAAI conference on artificial intelligence. Phoenix, Arizona, USA, pp 1266–1272
11. Liu Y, Nie L, Liu L, Rosenblum DS (2016) From action to activity: sensor-based activity recognition. Neurocomputing 181:108–115
12. Liu Y, Zhang L, Nie L, Yan Y, Rosenblum DS (2016) Fortune teller: predicting your career path. In: Proceedings of the thirtieth AAAI conference on artificial intelligence. Phoenix, Arizona, USA, pp 201–207
13. Liu Y, Zheng Y, Liang Y, Liu S, Rosenblum DS (2016) Urban water quality prediction based on multitask multi-view learning. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, IJCAI. New York, NY, USA, pp 2576–2581
14. Lu Y, Wei Y, Liu L, Zhong J, Sun L, Liu Y (2017) Towards unsupervised physical activity recognition using smartphone accelerometers. Multimed Tools Appl 76(8):10,701–10,719
15. Mairal J, Bach FR, Ponce J, Sapiro G, Zisserman A (2008) Discriminative learned dictionaries for local image analysis. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR 2008). Anchorage, Alaska, USA
16. Martinez A, Benavente R (1998) The AR face database. CVC Tech Rep 24
17. Nam S, Davies ME, Elad M, Gribonval R (2013) The cosparse analysis model and algorithms. Appl Computat Harmon Anal 34(1):30–56
18. Preotiuc-Pietro D, Liu Y, Hopkins D, Ungar LH (2017) Beyond binary labels: political ideology prediction of twitter users. In: Proceedings of the 55th annual meeting association computational linguistics, ACL. Vancouver, Canada, pp 729–740
19. Ravishankar S, Bresler Y (2013) Learning overcomplete sparsifying transforms for signal processing. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing. Vancouver, BC, Canada, pp 3088–3092
20. Ravishankar S, Bresler Y (2013) Learning sparsifying transforms. IEEE Trans Signal Process 61(5):1072–1086
21. Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Applicat 24(5):971–981
22. Rubinstein R, Elad M (2014) Dictionary learning for analysis-synthesis thresholding. IEEE Trans Signal Process 62(22):5962–5972
23. Rubinstein R, Peleg T, Elad M (2013) Analysis k-svd: a dictionary-learning algorithm for the analysis sparse model. IEEE Trans Signal Process 61(3):661–677
24. Schölkopf B, Platt J, Hofmann T (2006) Sparse representation for signal classification. In: Proceedings of the advances neural information processing systems. Vancouver, British Columbia, Canada, pp 609–616
25. Shekhar S, Patel VM, Chellappa R (2014) Analysis sparse coding models for image-based classification. In: Proceedings of the IEEE international conference on image processing. Paris, France, pp 5207–5211
26. Svetlana L, Cordelia S, Jean P (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 2. New York, USA, pp 2169–2178
27. Wang L, Zhang Y, Feng J (2005) On the euclidean distance of images. IEEE Trans Pattern Anal Mach Intell 27(8):1334–1339
28. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227

29. Xu Z, Guo Y, Guo J, Kong X (2015) Hybrid dictionary learning for JPEG steganalysis. In: Asia-pacific signal information process. Association annual summit conference, APSIPA 2015. Hong Kong, pp 711–714
30. Yang MH, Yang J (2012) Top-down visual saliency via joint crf and dictionary learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2296–2303
31. Zhang Q, Li B (2010) Discriminative k-svd for dictionary learning in face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. San Francisco, CA, pp 2691–2698
32. Zhang L, Yang M, Feng X (2011) Sparse representation or collaborative representation: which helps face recognition? In: Proceedings of the IEEE international conference on computer vision. Barcelona, Spain, pp 471–478



Qianyu Wang received her B.E. degree in English Intensive, Electronic and Information Engineering, Dalian University of Technology of China, in 2016. She is currently a master student in the School of Information and Communication Engineering, Dalian University of Technology. Her research interest is in dictionary learning and machine learning.

Yanqing Guo received the B.S. degree and Ph.D. degree in Electronic Engineering from Dalian University of Technology of China, in 2002 and 2009, respectively. He is currently an associate professor with Faculty of Electronic Information and Electrical Engineering Dalian University of Technology. His research interests include multimedia security and forensics, digital image processing and machine learning.

Jun Guo received the B.S. degree in electronics and information engineering and the M.S. degree in information and communication engineering from Dalian University of Technology of China, in 2013 and 2016, respectively. He is currently a Ph.D. candidate in Tsinghua-Berkeley Shenzhen Institute, Tsinghua University. His research interests include pattern recognition and machine learning. In particular, he focuses on dictionary learning, matrix factorization and their applications on multimedia and data processing.

Xiangwei Kong is a professor of Department of Electronic and Information Engineering, and vice director of Information Security Research Center of Dalian University of Technology, Dalian, China. She received B.E. and M.Sc. degree from Harbin Shipbuilding Engineering Institute in 1985 and 1988 and received Ph.D. degree from Dalian University of Technology, in 2003. Her research interests include multimedia security and forensics, digital image processing and pattern recognition.