A Dictionary Learning Approach for Classification: Separating the Particularity and the Commonality[*]

Shu Kong and Donghui Wang[**]

Dept. of Computer Science and Technology, Zhejiang University, Hangzhou, China
{aimerykong,dhwang}@zju.edu.cn

Abstract. Empirically, we find that, despite the class-specific features owned by the objects appearing in the images, the objects from different categories usually share some common patterns, which do not contribute to the discrimination of them. Concentrating on this observation and under the general dictionary learning (DL) framework, we propose a novel method to explicitly learn a common pattern pool (the commonality) and class-specific dictionaries (the particularity) for classification. We call our method DL-COPAR, which can learn the most compact and most discriminative class-specific dictionaries used for classification. The proposed DL-COPAR is extensively evaluated both on synthetic data and on benchmark image databases in comparison with existing DL-based classification methods. The experimental results demonstrate that DL-COPAR achieves very promising performances in various applications, such as face recognition, handwritten digit recognition, scene classification and object recognition.

Keywords: Dictionary Learning, Classification, Commonality, Particularity.

1 Introduction

Dictionary learning (DL), as a particular sparse signal model, has risen to prominence in recent years. It aims to learn a (overcomplete) dictionary in which only a few atoms can be linearly combined to well approximate a given signal. DL-based methods have achieved state-of-the-art performances in many application fields, such as image denoising [1] and image classification [2,3].

DL methods are originally designed to learn a dictionary which can faithfully represent signals, therefore they do not work well for classification tasks. To circumvent this problem, researchers recently develop several approaches to learn a classification-oriented dictionary. By exploring the label information, most DL-based classification methods learn such an adaptive dictionary mainly in two ways: either directly forcing the dictionary discriminative, or making the sparse coefficients discriminative (usually through simultaneously learning a classifier) to promote the discrimination power of the dictionary. For the first case, as an example, Ramirez *et al.* advocate learning class-specific sub-dictionaries for each class with a novel penalty term to make the sub-dictionaries incoherent [4]. Most methods belong to the latter case. Mairal *et al.* propose

[*] This work is financially supported by 973Program (No.2010CB327904) and Natural Science Foundations of China (No.61071218).

[**] Corresponding author.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 186–199, 2012.

a supervised DL method by embedding a logistic loss function to simultaneously learn a classifier [3]. Zhang and Li propose discriminative K-SVD method to achieve a desired dictionary which has good representation power while supporting optimal discrimination of the classes [5]. Furthermore, Jiang *et al.* add a label consistence term on K-SVD to make the sparse coefficients more discriminative, thus the discrimination power of the dictionary is further improved. Out of these two scenarios, Yang *et al.* propose Fisher discrimination DL method to simultaneously learn class-specific sub-dictionaries and to make the coefficients more discriminative through Fisher criterion [6].

Besides, in the empirical observation, despite the highly discriminative features owned by the objects from different categories, the images containing the objects usually share some common patterns, which do not contribute to the recognition of them. Such common patterns include, for example, the background of the objects in object categorization, the noses in expression recognition (the nose is almost motionless in various expressions), and so on. Under this observation, we can see it is not feasible to treat the (inter-category) atoms of the overall dictionary equally without discrimination. In the inspiring work of Ramirez *et al.* [4], a DL method is proposed by adding a structured inherence penalty term (DLSI) to learn C sub-dictionaries for C classes (each one corresponds to a specific class), and discarding the most coherent atoms among these sub-dictionaries as the shared features. These shared features will confuse the representation as a result of the coherent atoms from different sub-dictionaries can be used for representation interchangeably. Then DLSI uses the combination of all the sub-dictionaries as an overall dictionary for final classification. Even if some improvements are achieved by DLSI, the common patterns are still hidden in the sub-dictionaries. In the work of Wang *et al.* [7], an automatic group sparse coding (AutoGSC) method is proposed to discover the hidden shared data patterns with a common dictionary and C individual dictionaries for the C groups. Actually, AutoGSC is a clustering approach based on sparse coding and is not adapted for classification.

Inspired by the aforementioned works, we propose a novel DL-based approach for classification tasks, named **DL-COPAR** in this paper. Given the training data with the label information, we expect explicitly learning the class-specific feature dictionaries (the **particularity**) and the common pattern pool (the **commonality**). The particularity is most discriminative for classification despite its representation power, and the commonality is separated out to merely contribute the representation for the data from all classes. With the help of the commonality, the overall dictionary can be more compact and more discriminative for classification. To evaluate the proposed DL-COPAR, we extensively perform a series of experiments both on synthesis data and on benchmark image databases. The experimental results demonstrate DL-COPAR achieves very promising performances in various applications, such as face recognition, handwritten digit recognition, scene classification and object recognition.

2 Learning the Commonality and the Particularity

To derive our DL-COPAR, we first review the classical dictionary learning (DL) model. Suppose there are N training data (possibly from different categories) denoted by $\mathbf{x}_i \in \mathbb{R}^d$ $(i = 1, \dots, N)$, DL learns a dictionary $\mathbf{D} \in \mathbb{R}^{d \times K}$ from them by alternatively minimizing the objective function f over $\mathbf{D}$ and coefficient matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N] \in \mathbb{R}^{K \times N}$:

$$\{\mathbf{A}, \mathbf{D}\} = \underset{\mathbf{D},\mathbf{A}}{\operatorname{argmin}} \left\{ f \equiv \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda \|\mathbf{a}_i\|_1 \right\} \quad (1)$$
$$\text{s.t. } \|\mathbf{d}_k\|_2^2 = 1, \text{for } \forall k = 1, \ldots, K,$$

If $K > d$, then $\mathbf{D}$ is called an overcomplete dictionary. Suppose there are C classes and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_c, \ldots, \mathbf{X}_C] \in \mathbb{R}^{d \times N}$ is the dataset, wherein $\mathbf{X}_c \in \mathbb{R}^{d \times N_c}$ ($N = \sum_{c=1}^{C} N_c$) represent the data from the c^{th} class and signal $\mathbf{x}_i \in \mathbb{R}^d$ from this class is indexed by $i \in \mathcal{I}_c$. Although the learned dictionary $\mathbf{D}$ by Eq.(1) from $\mathbf{X}$ is adapted for representation of the data, it is not suitable for classifying them. To obtain a classification-oriented dictionary, an intuitive way is to learn C class-specific dictionaries $\mathbf{D}_c$'s for all C classes ($c = 1, \ldots, C$), then the reconstruction error can be used for classification [4,8]. Besides, as we observe in the real world, the class-specific dictionaries $\mathbf{D}_c$'s from different categories usually share some common patterns/atoms, which do not contribute to classification but are essential for reconstruction in DL. Thus these common/coherent atoms can be interchangeably used for representation of a query datum, by which way the classification performance can be compromised. For this reason, to improve classification performance, we explicitly separate the coherent atoms by learning the commonality $\mathbf{D}_{C+1}$, which provides the common bases for all categories. Denote the overall dictionary as $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_c, \ldots, \mathbf{D}_C, \mathbf{D}_{C+1}] \in \mathbb{R}^{d \times K}$, in which $K = \sum_{c=1}^{C+1} K_c$, $\mathbf{D}_c \in \mathbb{R}^{d \times K_c}$ stands for the particularity of the c^{th} class, and $\mathbf{D}_{C+1} \in \mathbb{R}^{d \times K_{C+1}}$ is the commonality. $\mathbf{I}$ is the identity matrix with appropriate size.

First of all, a learned dictionary $\mathbf{D}$ ought to well represent every sample $\mathbf{x}_i$, *i.e.* $\mathbf{x}_i \approx \mathbf{D}\mathbf{a}_i$, where the efficient $\mathbf{a}_i = [\boldsymbol{\theta}_i^{(1)}; \ldots; \boldsymbol{\theta}_i^{(C)}; \boldsymbol{\theta}_i^{(C+1)}] \in \mathbb{R}^K$ and $\boldsymbol{\theta}_i^{(c)} \in \mathbb{R}^{K_c}$ is the part corresponding to the sub-dictionary $\mathbf{D}_c$. Despite of the overall dictionary, it is also expected that the sample from the c^{th} class can be well represented by the cooperative efforts of the c^{th} particularity $\mathbf{D}_c$ and the commonality $\mathbf{D}_{C+1}$. Therefore, we renew the objective function f:

$$f \equiv \sum_{c=1}^{C} \sum_{i \in \mathcal{I}_c} \left\{ \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda \|\mathbf{a}_i\|_1 + \|\mathbf{x}_i - \mathbf{D}_c \boldsymbol{\theta}_i^{(c)} - \mathbf{D}_{C+1} \boldsymbol{\theta}_i^{(C+1)}\|_2^2 \right\}, \quad (2)$$

where $\boldsymbol{\theta}_i^{(C)}$ and $\boldsymbol{\theta}_i^{(C+1)}$ are the corresponding coefficients of the two sub-dictionaries. Actually, the last term of Eq.(2) is the same as the objective formulation of [7] by ignoring the sparse penalty term. For mathematical brevity, we introduce a selection operator $\mathbf{Q}_c = [\mathbf{q}_1^c, \ldots, \mathbf{q}_j^c, \ldots, \mathbf{q}_{K_c}^c] \in \mathbb{R}^{K \times K_c}$, in which the j^{th} column of $\mathbf{Q}_c$ is of the form:

$$\mathbf{q}_j^c = [\, \underbrace{0, \ldots, 0}_{\sum_{m=1}^{c-1} K_m}, \underbrace{\overbrace{0, \ldots, 0}^{j-1}, 1, 0, \ldots, 0}_{K_c}, \underbrace{0, \ldots, 0}_{\sum_{m=c+1}^{C+1} K_m} \,]^T. \quad (3)$$

Therefore, we have $\mathbf{Q}_c^T \mathbf{Q}_c = \mathbf{I}$, $\mathbf{D}_c = \mathbf{D}\mathbf{Q}_c$ and $\boldsymbol{\theta}_i^{(c)} = \mathbf{Q}_c^T \mathbf{a}_i \in \mathbb{R}^{K_c}$. Let $\tilde{\mathbf{Q}}_c = [\mathbf{Q}_c, \mathbf{Q}_{C+1}]$, and we have the coefficient corresponding to the particularity and the commonality $\begin{pmatrix}\boldsymbol{\theta}_i^{(c)}\\ \boldsymbol{\theta}_i^{(C+1)}\end{pmatrix} = \tilde{\mathbf{Q}}_c^T\mathbf{a}_i \in \mathbb{R}^{K_c+K_{C+1}}$. By introducing a terse denotation $\phi(\mathbf{A}_c) = \sum_{i=1}^{N_c} \|\mathbf{a}_i^c\|_1$ for $\mathbf{A}_c = [\mathbf{a}_1^c, \ldots, \mathbf{a}_{N_c}^c]$, Eq.(2) can be rewritten as:

$$f \equiv \sum_{c=1}^{C} \{ \|\mathbf{X}_c - \mathbf{D}\mathbf{A}_c\|_F^2 + \lambda\phi(\mathbf{A}_c) + \|\mathbf{X}_c - \mathbf{D}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c\|_F^2. \} \tag{4}$$

In this way, we first guarantee the representation power of the overall dictionary, and then require the particularity, in conjunction with the commonality, to well represent the data of this category. However, merely doing this is not sufficient to learn a discriminative dictionary, because other class-specific dictionaries may share similar bases with that of the c^{th} one, *i.e.* the atoms from different class-specific dictionaries can still be coherent and thus be used interchangeably for representing the query data. To circumvent this problem, we force the coefficients, except the parts corresponding to the c^{th} particularity and the commonality, to be zero. Mathematically, denote $\mathbf{Q}_{/c} = [\mathbf{Q}_1, \ldots, \mathbf{Q}_{c-1}, \mathbf{Q}_{c+1}, \ldots, \mathbf{Q}_C, \mathbf{Q}_{C+1}]$ and $\tilde{\mathbf{Q}}_{/c} = [\mathbf{Q}_1, \ldots, \mathbf{Q}_{c-1}, \mathbf{Q}_{c+1}, \ldots, \mathbf{Q}_C]$, we force $\|\tilde{\mathbf{Q}}_{/c}^T\mathbf{A}_c\|_F^2 = 0$. Thus, we add a new term to the objective function:

$$f \equiv \sum_{c=1}^{C} \left\{ \|\mathbf{X}_c - \mathbf{D}\mathbf{A}_c\|_F^2 + \|\tilde{\mathbf{Q}}_{/c}^T\mathbf{A}_c\|_F^2 + \lambda\phi(\mathbf{A}_c) + \|\mathbf{X}_c - \mathbf{D}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c\|_F^2 \right\} \tag{5}$$

It is worth noting that Eq.(5) can fail in capturing the common patterns. For example, the bases of the real common patterns may appear in several particularities, which makes the learned particularities redundant and less discriminative. Therefore, it is desired to drive the common patterns to the commonality and preserve the class-specific features in the particularity. For this purpose, we add an incoherence term $\mathcal{Q}(\mathbf{D}_i, \mathbf{D}_j) = \|\mathbf{D}_i^T\mathbf{D}_j\|_F^2$ to the objective function. This penalty term has been used among the class-specific sub-dictionaries in [4], but we also consider the incoherence of the commonality with the particularities. Then we derive the objective function of the proposed DL-COPAR:

$$f \equiv \sum_{c=1}^{C} \left\{ \begin{matrix} \|\mathbf{X}_c - \mathbf{D}\mathbf{A}_c\|_F^2 + \|\tilde{\mathbf{Q}}_{/c}^T\mathbf{A}_c\|_F^2 \\ \|\mathbf{X}_c - \mathbf{D}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c\|_F^2 + \lambda\phi(\mathbf{A}_c) \end{matrix} \right\} + \eta \sum_{c=1}^{C+1} \sum_{\substack{j=1 \\ j\neq c}}^{C+1} \mathcal{Q}(\mathbf{D}_c, \mathbf{D}_j) \tag{6}$$

3 Optimization Step

At the first sight, the objective function Eq.(6) is complex to solve, but we show it can be easily optimized through an alternative optimization process.

3.1 Fix $\mathbf{D}$ to Update $\mathbf{A}_c$

When fixing $\mathbf{D}$ to update $\mathbf{A}_c$, we omit the terms independent of $\mathbf{A}_c$ from Eq.(6):

$$\begin{aligned} f &\equiv \sum_{c=1}^{C} \left\{ \|\mathbf{X}_c - \mathbf{D}\mathbf{A}_c\|_F^2 + \|\tilde{\mathbf{Q}}_{/c}^T\mathbf{A}_c\|_F^2 + \lambda\phi(\mathbf{A}_c) + \|\mathbf{X}_c - \mathbf{D}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c\|_F^2 \right\} \\ &= \sum_{c=1}^{C} \left\{ \left\| \begin{pmatrix} \mathbf{X}_c \\ \mathbf{X}_c \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \mathbf{D}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T \\ \tilde{\mathbf{Q}}_{/c}^T \end{pmatrix} \mathbf{A}_c \right\|_F^2 + \lambda\phi(\mathbf{A}_c) \right\}. \end{aligned} \tag{7}$$

Proposition 1. *Denote* $\bar{\mathbf{X}}_c = \begin{pmatrix} \mathbf{X}_c \\ \mathbf{X}_c \\ \mathbf{0} \end{pmatrix}$ *and* $\bar{\mathbf{D}} = [\bar{\mathbf{d}}_1, \ldots, \bar{\mathbf{d}}_k, \ldots, \bar{\mathbf{d}}_K]$*, where* $\bar{\mathbf{d}}_k$ *is the* ℓ_2*-norm unitized vector of the* k^{th} *column in* $\begin{pmatrix} \mathbf{D} \\ \mathbf{D}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T \\ \tilde{\mathbf{Q}}_{/c}^T \end{pmatrix}$*, then when fixing* $\mathbf{D}$ *to minimize Eq.(6) over* $\mathbf{A}_c$*, the minimum is reached when* $\mathbf{A}_c$ *equals to* $\frac{1}{\sqrt{2}}\bar{\mathbf{A}}$*, wherein* $\bar{\mathbf{A}}$ *is the result of:*

$$\bar{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmin}} \|\bar{\mathbf{X}}_c - \bar{\mathbf{D}}\mathbf{A}\|_F^2 + \frac{\lambda}{\sqrt{2}}\phi(\mathbf{A}),$$

Proof. Denote $\mathbf{D}' = \begin{pmatrix} \mathbf{D} \\ \mathbf{D}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T \\ \tilde{\mathbf{Q}}_{/c}^T \end{pmatrix}$, through simple derivations, it is easy to show the Euclidean length of each column of $\mathbf{D}'$ is $\sqrt{2}$, then

$$\begin{aligned} \mathbf{A}_c &= \underset{\mathbf{A}_c}{\operatorname{argmin}} \left\{ \|\bar{\mathbf{X}}_c - \tfrac{1}{\sqrt{2}}\mathbf{D}'\sqrt{2}\mathbf{A}_c\|_F^2 + \lambda\phi(\mathbf{A}_c) \right\} \\ &= \underset{\mathbf{A}_c}{\operatorname{argmin}} \left\{ \|\bar{\mathbf{X}}_c - \bar{\mathbf{D}}\sqrt{2}\mathbf{A}_c\|_F^2 + \tfrac{\lambda}{\sqrt{2}}\phi(\sqrt{2}\mathbf{A}_c) \right\} \end{aligned} \tag{8}$$

Proposition 1 indicates that to update the coefficients $\mathbf{A}_c$ with fixed $\mathbf{D}$, we can simply solve a LASSO problem. Through this paper, we adopt the feature-sign search algorithm [9] to solve this LASSO problem.

3.2 Fix $\mathbf{A}_c$ to Update $\mathbf{D}$

To update the overall dictionary $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_C, \mathbf{D}_{C+1}]$, we turn to an iterative approach, *i.e.* updating $\mathbf{D}_c$ by fixing all the other $\mathbf{D}_i$'s. Despite the c^{th} class-specific dictionary, the common pattern pool also contributes to fitting the signals from the c^{th} class. Thus there are differences in optimizing $\mathbf{D}_{C+1}$ and $\mathbf{D}_c$. We elaborate the optimization steps as below.

Update the Particularity $\mathbf{D}_c$. Without loss of generality, we concentrate on the optimization of the c^{th} class-specific dictionary $\mathbf{D}_c$ by fixing the other $\mathbf{D}_i$'s. Specifically, we denote $\mathbf{A}_c^{(i)} = \mathbf{Q}_i^T\mathbf{A}_c$ for $i = 1, \ldots, C+1$. By dropping the unrelated terms, we update the c^{th} particularity $\mathbf{D}_c$ as below:

$$\mathbf{D}_c = \underset{\mathbf{D}_c}{\operatorname{argmin}} \left\{ \begin{array}{l} \|\mathbf{X}_c - \sum\limits_{\substack{i=1 \\ i \neq c}}^{C+1} \mathbf{D}_i\mathbf{A}_c^{(i)} - \mathbf{D}_c\mathbf{A}_c^{(c)}\|_F^2 + \\ \|\mathbf{X}_c - \mathbf{D}_{C+1}\mathbf{A}_c^{(C+1)} - \mathbf{D}_c\mathbf{A}_c^{(c)}\|_F^2 + 2\eta \sum\limits_{\substack{i=1 \\ i \neq c}}^{C+1} \mathcal{Q}(\mathbf{D}_c, \mathbf{D}_i) \end{array} \right\} \tag{9}$$

Let $\bar{\mathbf{X}}_c = \mathbf{X}_c - \mathbf{D}_{C+1}\mathbf{A}_c^{(C+1)}$, $\bar{\mathbf{Y}}_c = \mathbf{X}_c - \sum_{i=1, i \neq c}^{C+1} \mathbf{D}_i\mathbf{A}_c^{(i)}$ and $\mathbf{B} = \mathbf{D}\mathbf{Q}_{/c}$, then we have:

$$\mathbf{D}_c = \underset{\mathbf{D}_c}{\operatorname{argmin}} \|\bar{\mathbf{Y}}_c - \mathbf{D}_c\mathbf{A}_c^{(c)}\|_F^2 + \|\bar{\mathbf{X}}_c - \mathbf{D}_c\mathbf{A}_c^{(c)}\|_F^2 + 2\eta\|\mathbf{D}_c^T\mathbf{B}\|_F^2. \tag{10}$$

We propose to update $\mathbf{D}_c = [\mathbf{d}_c^1, \ldots, \mathbf{d}_c^{K_c}]$ atom by atom, *i.e.* updating $\mathbf{d}_c^k$ while fixing other columns. Specifically, denote $\mathbf{A}_c^{(c)} = [\mathbf{a}_{(1)}; \ldots; \mathbf{a}_{(K_c)}] \in \mathbb{R}^{K_c \times N_c}$, wherein $\mathbf{a}_{(k)} \in \mathbb{R}^{1 \times N_c}$ is the k^{th} row of $\mathbf{A}_c^{(c)}$. Let $\hat{\mathbf{Y}}_c = \bar{\mathbf{Y}}_c - \sum_{j \neq k} \mathbf{d}_c^j \mathbf{a}_{(j)}$ and $\hat{\mathbf{X}}_c = \bar{\mathbf{X}}_c - \sum_{j \neq k} \mathbf{d}_c^j \mathbf{a}_{(j)}$, then we get:

$$\mathbf{d}_c^k = \underset{\mathbf{d}_c^k}{\operatorname{argmin}} \left\{ g(\mathbf{d}_c^k) \equiv \|\hat{\mathbf{Y}}_c - \mathbf{d}_c^k \mathbf{a}_{(k)}\|_F^2 + \|\hat{\mathbf{X}}_c - \mathbf{d}_c^k \mathbf{a}_{(k)}\|_F^2 + 2\eta \|\mathbf{d}_c^{k\,T} \mathbf{B}\|_F^2 \right\}. \quad (11)$$

Let the first derivative of $g(\mathbf{d}_c^k)$ w.r.t. $\mathbf{d}_c^k$ equal zero, *i.e.* $\partial g(\mathbf{d}_c^k)/\partial \mathbf{d}_c^k = 0$, then we obtain the updated $\mathbf{d}_c^k$ as below:

$$\mathbf{d}_c^k = \frac{1}{2}(\|\mathbf{a}_{(k)}\|_2^2 \mathbf{I} + \eta \mathbf{B}\mathbf{B}^T)^{-1}(\hat{\mathbf{Y}}_c + \hat{\mathbf{X}}_c)\mathbf{a}_{(k)}^T. \quad (12)$$

Note as an atom of dictionary, it ought to be unitized, *i.e.* $\hat{\mathbf{d}}_c^k = \mathbf{d}_c^k / \|\mathbf{d}_c^k\|_2$. Along with this, the corresponding coefficient should be multiplied $\|\mathbf{d}_c^k\|_2$, *i.e.* $\hat{\mathbf{a}}_{(k)} = \|\mathbf{d}_c^k\|_2 \mathbf{a}_{(k)}$.

Update the Shared Feature Pool $\mathbf{D}_{C+1}$. Denote $\mathbf{B} = \mathbf{D}\mathbf{Q}_{/C+1}$, *i.e.* $\mathbf{B} = [\mathbf{D}_1, \ldots, \mathbf{D}_C]$, by dropping the unrelated terms, we update $\mathbf{D}_{C+1}$ as below:

$$\begin{aligned} \mathbf{D}_{C+1} = \underset{\mathbf{D}_{C+1}}{\operatorname{argmin}} \sum_{c=1}^{C} \Big\{ &\|\mathbf{X}_c - \sum_{i=1}^{C} \mathbf{D}_i \mathbf{A}_c^{(i)} - \mathbf{D}_{C+1}\mathbf{A}_{C+1}^{(C+1)}\|_F^2 + \\ &\|\mathbf{X}_c - \mathbf{D}_c \mathbf{A}_c^{(c)} - \mathbf{D}_{C+1}\mathbf{A}_{C+1}^{(C+1)}\|_F^2 \Big\} + 2\eta \|\mathbf{D}_{C+1}^T \mathbf{B}\|_F^2. \end{aligned} \quad (13)$$

Denote $\bar{\mathbf{Y}}_c = \mathbf{X}_c - \sum_{i=1}^{C} \mathbf{D}_i \mathbf{A}_c^{(i)}$ and $\bar{\mathbf{X}}_c = \mathbf{X}_c - \mathbf{D}_c \mathbf{A}_c^{(c)}$ (note $\bar{\mathbf{X}}_c$ and $\bar{\mathbf{Y}}_c$ here are different from the ones in the optimization of $\mathbf{D}_c$), then we have:

$$\mathbf{D}_{C+1} = \underset{\mathbf{D}_{C+1}}{\operatorname{argmin}} \|\bar{\mathbf{Y}} - \mathbf{D}_{C+1}\mathbf{A}^{(C+1)}\|_F^2 + \|\bar{\mathbf{X}} - \mathbf{D}_{C+1}\mathbf{A}^{(C+1)}\|_F^2 + 2\eta \|\mathbf{D}_{C+1}^T \mathbf{B}\|_F^2, \quad (14)$$

where $\mathbf{A}^{(C+1)} = [\mathbf{A}_1^{(C+1)}, \ldots, \mathbf{A}_C^{(C+1)}]$, $\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1, \ldots, \bar{\mathbf{X}}_C]$, and $\bar{\mathbf{Y}} = [\bar{\mathbf{Y}}_1, \ldots, \bar{\mathbf{Y}}_C]$. As well, we choose to update $\mathbf{D}_{C+1}$ atom by atom, and the k^{th} column is updated by:

$$\mathbf{d}_{C+1}^k = \frac{1}{2}(\|\mathbf{a}_{(k)}\|_2^2 \mathbf{I} + \eta \mathbf{B}\mathbf{B}^T)^{-1}(\hat{\mathbf{X}} + \hat{\mathbf{Y}})\mathbf{a}_{(k)}^T, \quad (15)$$

where $\mathbf{A}^{(C+1)} = [\mathbf{a}_{(1)}; \ldots; \mathbf{a}_{(K_{C+1})}] \in \mathbb{R}^{K_{C+1} \times N}$, $\hat{\mathbf{Y}} = \bar{\mathbf{Y}} - \sum_{j \neq k} \mathbf{d}_{C+1}^j \mathbf{a}_{(j)}$, and $\hat{\mathbf{X}} = \bar{\mathbf{X}} - \sum_{j \neq k} \mathbf{d}_{C+1}^j \mathbf{a}_{(j)}$. Similarly, we unitize $\mathbf{d}_{C+1}^k$ to get the unit-length atom $\hat{\mathbf{d}}_{C+1}^k$, with scaled coefficients $\hat{\mathbf{a}}_{(k)} = \|\mathbf{d}_{C+1}^k\|_2 \mathbf{a}_{(k)}$.

The overall algorithm is summarized in Algorithm 1. Note that the value of the objective function decreases at each iteration, therefore the algorithm converges.

Algorithm 1. Learning the Commonality and Particularity

```
Require: training dataset $[\mathbf{X}_1, \ldots, \mathbf{X}_C]$ with labels, the size of the $C+1$ desired sub-dictionaries $K_c$'s, and $\lambda$;
Ensure: $\|\mathbf{d}_i\|_2 = 1$, for $\forall i = 1, \ldots, K$;
initialize $\mathbf{D}_c$ for $c = 1, \ldots, C, C+1$;
while stop criterion is not reached do
    update the coefficients $\mathbf{A}_c$ by solving LASSO problem Eq.(8);
    update $\mathbf{D}_c$ atom-by-atom through Eq.(12) with unitization process and correspondingly scale the related coefficients;
    update $\mathbf{D}_{C+1}$ atom-by-atom through Eq.(15) with unitization process and correspondingly scale the related coefficients;
end while
return the learned dictionaries $\mathbf{D}_1, \ldots, \mathbf{D}_C, \mathbf{D}_{C+1}$.
```

4 Experimental Validation

In this section, we perform a series of experiments to evaluate the proposed DL-COPAR. First, a synthetic dataset is used to demonstrate the powerfulness of our method in learning the commonality and particularity. Then, we conduct experiments to compare our method with some state-of-the-art approaches on four public available benchmarks for four real-world recognition tasks respectively, *i.e.* face, hand-written digit, scene, and object recognition.

4.1 Experimental Setup

We employ K-SVD algorithm [10] to initialize the particularities and the commonality. In detail, to initialize the c^{th} class-specific dictionary, we perform K-SVD on the data from the c^{th} category, and to initialize the common pattern pool, we perform K-SVD on the whole training set. Note that we adopt feature-sign search algorithm [9] for the sparse coding problem throughout our experiments.

In classification stage, we propose to adopt two reconstruction error based classification schemes and the linear SVM classifier for different classification tasks. The two reconstruction error based classification schemes are used for the face and hand-written digit recognition, one is a global classifier (GC) and the other is a local classifier (LC). Note throughout our experiments, the sparsity parameter γ as below is tuned to make about 20% elements of the coefficient be nonzero.

GC For a testing sample $\mathbf{y}$, we code it over the overall learned dictionary $\mathbf{D}$, $\mathbf{a} = \operatorname{argmin}_{\mathbf{a}} \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 + \gamma\|\mathbf{a}\|_1$, where γ is a constant. Denote $\mathbf{a} = [\boldsymbol{\theta}^{(1)}; \ldots; \boldsymbol{\theta}^{(C)}; \boldsymbol{\theta}^{(C+1)}]$, where $\boldsymbol{\theta}^{(c)}$ is the coefficient vector associated with sub-dictionary $\mathbf{D}_c$. Then the reconstruction error by class c is $e_c = \|\mathbf{y} - \hat{\mathbf{D}}_c \hat{\boldsymbol{\theta}}^{(c)}\|_2^2$, where $\hat{\mathbf{D}}_c = [\mathbf{D}_c, \mathbf{D}_{C+1}]$ and $\hat{\boldsymbol{\theta}}^{(c)} = [\boldsymbol{\theta}^{(c)}; \boldsymbol{\theta}^{(C+1)}]$. Finally, the predicted label $\hat{c}$ is calculated by $\hat{c} = \operatorname{argmin}_c e_c$.

LC For a testing sample $\mathbf{y}$, we directly calculate C reconstruction error for the C classes: $e_c = \min_{\hat{\mathbf{a}}_c} \|\mathbf{y} - \hat{\mathbf{D}}_c \hat{\mathbf{a}}_c\|_2^2 + \gamma\|\hat{\mathbf{a}}_c\|_1$, where $\hat{\mathbf{D}}_c = [\mathbf{D}_c, \mathbf{D}_{C+1}]$. The final identity of $\mathbf{y}$ is $\hat{c} = \operatorname{argmin}_c e_c$.

The linear SVM is used for scene and object recognition with spatial pyramid matching (SPM) framework [11]. But in the dictionary learning stage, we adopt the proposed DL-COPAR. In detail, we first extract SIFT descriptors (128 in length) [12] from 16×16 pixel patches, which are densely sampled from each image on a dense grid with 8-pixels stepsize. Over the extracted SIFT descriptors from different categories, we use the proposed DL-COPAR to train an overall dictionary $\mathbf{D} \in \mathbb{R}^{128\times K}$ concatenated by the particularities and the commonality. Then we encode the descriptors of the image over the learned dictionary. By partitioning the image into into $2^l \times 2^l$ segments in different scales $l = 0, 1, 2$, we use max pooling technique to pool the coefficients over each segment into a single vector. Thus, there are $1 + 2 \times 2 + 4 \times 4 = 21$ pooled vectors, which are concatenated into the $21K$-dimensional vector for final representation. Note that different from pooling process in the existing SPM-based methods, we only use the coefficients which correspond to the particularities by discarding the ones to the commonality. Finally, we concatenate all the intercepted pooled vectors as the representation of the image. In all the four recognition experiments, we run 10 times of each experiment and report the average recognition rate.

4.2 Synthetic Example

The synthetic dataset constitutes two classes of gray-scale images of size 30×30. Both categories have three common basis images denoted by $\mathbf{F}_i^{\mathbf{S}}$ for $i = 1, 2, 3$, as illustrated by right panel of Fig. 1, where we use dark color to represent value zero and white color to represent 255. Category 1 has four individual basis images denoted by $\mathbf{F}_{1i}^{\mathbf{I}}$ for $i = 1, \ldots, 4$, as shown in left panel of Fig. 1. Similarly, the four basis images of category 2 are denoted by $\mathbf{F}_{2i}^{\mathbf{I}}$ for $i = 1, \ldots, 4$ in Fig. 1. The set of images used in our experiment are generated by adding Gaussian noise with zero mean and 0.01 variance. Fig. 2 displays several examples from the dataset, the top row for the first class and the bottom row for the second class. Note a similar synthetic dataset is used in [7], but ours is different from it in two ways. We use Gaussian noise instead of uniform random noise, and noises are added to both the dark and the white region rather than merely to the dark region in [7]. Therefore, our synthetic data set is more difficult to separate the common patterns from the individual features.

We compare the learned dictionaries with that learned by two closely related DL-based classification methods. The methods used for comparison include Fisher Discriminative Dictionary Learning method (FDDL) [6] and the method based on DL with structured incoherence (DLSI) [4]. As a supervised DL method, FDDL learns class-specific dictionaries for each class and makes them most discriminative through Fisher criteria. Fig. 3 shows the learned dictionary by FDDL, one row for one class. It is clear that the dictionaries learned by FDDL are much discriminative and each class-specific dictionary can faithfully represent the signals from the corresponding class. However, even though FDDL succeeds in learning most discriminative dictionaries, it cannot separate the common patterns and the individual features. Thus, the dictionaries can be too redundant for classification. DLSI can learn the common features despite the class-specific dictionaries, as Fig. 4 shows. Actually, DLSI learns class-specific sub-dictionaries, and then select the most coherent atoms among different individual dictionaries as the common features. From this figure, we can see the learned individual

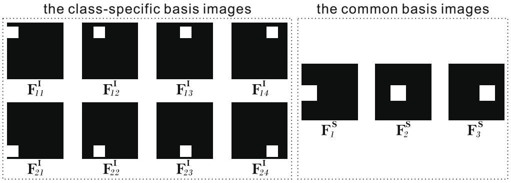
the class-specific basis images
the common basis images
$\mathbf{F}^{\mathbf{I}}_{11}$ $\mathbf{F}^{\mathbf{I}}_{12}$ $\mathbf{F}^{\mathbf{I}}_{13}$ $\mathbf{F}^{\mathbf{I}}_{14}$
$\mathbf{F}^{\mathbf{I}}_{21}$ $\mathbf{F}^{\mathbf{I}}_{22}$ $\mathbf{F}^{\mathbf{I}}_{23}$ $\mathbf{F}^{\mathbf{I}}_{24}$
$\mathbf{F}^{\mathbf{S}}_{1}$ $\mathbf{F}^{\mathbf{S}}_{2}$ $\mathbf{F}^{\mathbf{S}}_{3}$

Fig. 1. The real class-specific features and the common patterns

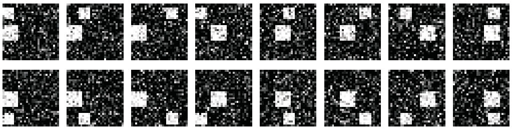

Fig. 2. Examples of the synthetic dataset

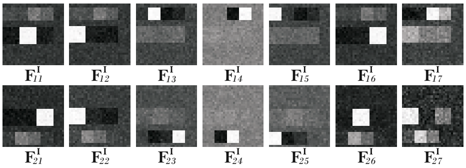
$\mathbf{F}^{\mathbf{I}}_{11}$ $\mathbf{F}^{\mathbf{I}}_{12}$ $\mathbf{F}^{\mathbf{I}}_{13}$ $\mathbf{F}^{\mathbf{I}}_{14}$ $\mathbf{F}^{\mathbf{I}}_{15}$ $\mathbf{F}^{\mathbf{I}}_{16}$ $\mathbf{F}^{\mathbf{I}}_{17}$
$\mathbf{F}^{\mathbf{I}}_{21}$ $\mathbf{F}^{\mathbf{I}}_{22}$ $\mathbf{F}^{\mathbf{I}}_{23}$ $\mathbf{F}^{\mathbf{I}}_{24}$ $\mathbf{F}^{\mathbf{I}}_{25}$ $\mathbf{F}^{\mathbf{I}}_{26}$ $\mathbf{F}^{\mathbf{I}}_{27}$

Fig. 3. The class-specific dictionaries learned by FDDL [6]

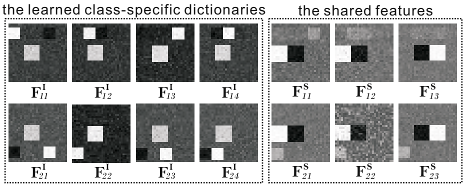
the learned class-specific dictionaries
the shared features
$\mathbf{F}^{\mathbf{I}}_{11}$ $\mathbf{F}^{\mathbf{I}}_{12}$ $\mathbf{F}^{\mathbf{I}}_{13}$ $\mathbf{F}^{\mathbf{I}}_{14}$
$\mathbf{F}^{\mathbf{S}}_{11}$ $\mathbf{F}^{\mathbf{S}}_{12}$ $\mathbf{F}^{\mathbf{S}}_{13}$
$\mathbf{F}^{\mathbf{I}}_{21}$ $\mathbf{F}^{\mathbf{I}}_{22}$ $\mathbf{F}^{\mathbf{I}}_{23}$ $\mathbf{F}^{\mathbf{I}}_{24}$
$\mathbf{F}^{\mathbf{S}}_{21}$ $\mathbf{F}^{\mathbf{S}}_{22}$ $\mathbf{F}^{\mathbf{S}}_{23}$

Fig. 4. The class-specific dictionaries and the shared features learned by DLSI [4]

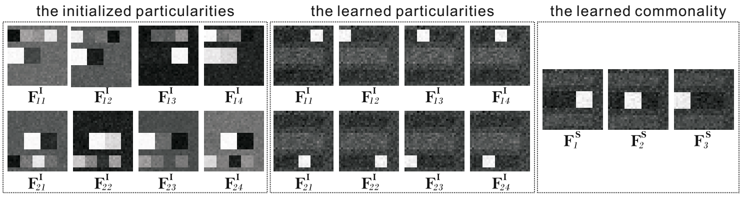
the initialized particularities
the learned particularities
the learned commonality
$\mathbf{F}^{\mathbf{I}}_{11}$ $\mathbf{F}^{\mathbf{I}}_{12}$ $\mathbf{F}^{\mathbf{I}}_{13}$ $\mathbf{F}^{\mathbf{I}}_{14}$
$\mathbf{F}^{\mathbf{I}}_{21}$ $\mathbf{F}^{\mathbf{I}}_{22}$ $\mathbf{F}^{\mathbf{I}}_{23}$ $\mathbf{F}^{\mathbf{I}}_{24}$
$\mathbf{F}^{\mathbf{I}}_{11}$ $\mathbf{F}^{\mathbf{I}}_{12}$ $\mathbf{F}^{\mathbf{I}}_{13}$ $\mathbf{F}^{\mathbf{I}}_{14}$
$\mathbf{F}^{\mathbf{I}}_{21}$ $\mathbf{F}^{\mathbf{I}}_{22}$ $\mathbf{F}^{\mathbf{I}}_{23}$ $\mathbf{F}^{\mathbf{I}}_{24}$
$\mathbf{F}^{\mathbf{S}}_{1}$ $\mathbf{F}^{\mathbf{S}}_{2}$ $\mathbf{F}^{\mathbf{S}}_{3}$

Fig. 5. The initialized particularities and the learned particularities and commonality by the proposed DL-COPAR

dictionaries are also mixed with the common features, as well the separated common features are too redundant.

On the contrary, as expected, the proposed DL-COPAR can correctly separate the common features and individual features, as shown in Fig. 5. As a comparison, we also plot the initialized particularities by K-SVD. There is no doubt that any signals can be

Table 1. The recognition rates (%) of various methods on Extended YaleB face database

Method	SRC	D-KSVD	LC-KSVD	DLSI	**FDDL**	*k*NN	SVM	DL-COPAR (LC) **DL-COPAR (GC)**
Accuracy	97.2	94.1	96.7	96.5	**97.9**	82.5	96.2	96.9 **98.3**

well represented by the corresponding class-specific dictionary and the shared pattern pool. This phenomenon reveals that DL-COPAR can learn the most compact and most discriminative dictionaries for classification.

4.3 Face Recognition

Extended YaleB [13] database contains 2,414 frontal face images of 38 persons, about 68 images for each individual. It is challenging due to varying illumination conditions and expressions, and the original images are cropped to 192 × 168 pixels. We use random faces [14,15,5] as the feature descriptors, which are obtained by projecting a face image onto a 504-dimensional vector with a randomly generated matrix from a zero-mean normal distribution. We randomly select half of the images (about 32 images per individual) for training and the rest for testing. The learned dictionary consists of 570 atoms, which correspond to an average of 15 atoms for each person and the last 5 ones as the common features. We compare the performance of our DL-COPAR with that of D-KSVD [5], SRC [14], DLSI [4], LC-KSVD [15] and FDDL [6].

Direct comparison is shown in Table 1. We can see that D-KSVD, which only uses coding coefficients for classification, does not work well on this dataset. LC-KSVD, which employs the label information to improve the discrimination of the learned overall dictionary and also uses the coefficients for the final classification, achieves better classification performance than D-KSVD, but still no better than that of SRC. DLSI aims to make the class-specific dictionaries incoherent and thereby facilitates the discrimination of the dictionaries. It employs the reconstruction error for classification, and achieves no better results than that of SRC. FDDL utilizes Fisher criterion to make the coefficients more discriminative, thus improves the discrimination of the individual dictionaries. It has achieved state-of-the-art classification performance on this database, with the accuracy higher than that of SRC. However, our DL-COPAR with GC achieves the best classification performance, about 0.4% higher than that of FDDL. But we can see if DL-COPAR use LC for classification, it cannot produce satisfactory outcome. The reason is that the size of the class-specific dictionary is too small to faithfully represent the data, thus the collaboration of these sub-dictionaries is of crucial significance.

4.4 Hand-Written Digit Recognition

USPS [1] is a widely used handwritten digit data set with 7, 291 training and 2, 007 testing images. We compare the proposed DL-COPAR with several methods including the reconstructive DL method with linear and bilinear classifier models (denoted by REC-L and REC-BL) reported in [3], the supervised DL method with generative training and

[1] http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html

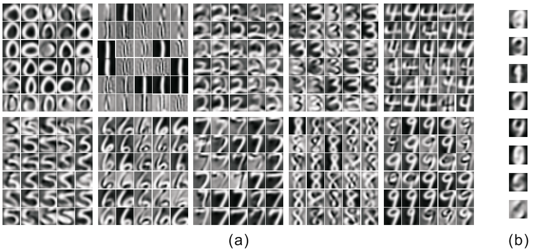

Fig. 6. The learned particularity (a) and commonalities (b) by DL-COPAR on USPS database

Table 2. Error rates (%) of various methods on USPS data set

Method	SRSC	REC-L REC-BL	SDL-G **SDL-D**	DLSI	FDDL	kNN	SVM	**DL-COPAR (LC)** DL-COPAR (GC)
Error Rate	6.05	6.83 4.38	6.67 **3.54**	3.98	3.69	5.2	4.2	**3.61** 4.70

discriminative training (denoted by SDL-G and SDL-D) reported in [3], sparse representation for signal classification (denoted by SRSC) by [16], DLSI [4] and FDDL [6]. Additionally, some results of problem-specific methods (*i.e.* the kNN method and SVM with a Gaussian kernel) reported in [4] are also listed. The original images of 16×16 resolution are directly used here.

Direct comparison is shown in Table 2. These results are originally reported or taken from the paper [6]. As we can see from the table, our DL-COPAR with LC classification scheme outperforms all the other methods except for SDL-D. Actually, the optimization of SDL-D is much more complex than that of DL-COPAR, and it uses much more information in the dictionary learning and classification process, such as a learned classifier of coding coefficients, the sparsity of coefficients and the reconstruction error. It is worth noting that DL-COPAR only trains 30 codewords for each class and 8 atoms as the shared commonality, whereas FDDL learns 90 atoms of dictionary for each digit even though it achieves very close result with that of DL-COPAR. Fig. 6 illustrates the learned particularities and the commonality by DL-COPAR, respectively.

4.5 Scene Recognition

We also try our DL-COPAR for scene classification task on 15-scene dataset, which is compiled by several researchers [22,23,17]. This dataset contains totally 4484 images falling into 15 categories, with the number of images per category ranging from 200 to 400. Following the common setup on this database [11,17], we randomly select 100 images per category as the training set and use the rest for testing. An overall 1024-visual-words dictionary is constituted. As for DL-COPAR, a 60-visual-word particularity is learned for each category, and the commonality consists of 124 shared pattern bases. Thus the size of the concatenated dictionary is $15 \times 60 + 124 = 1024$. Note that the size of the final representation of each image is $21 \times 15 \times 60 = 18900$, which is less

Table 3. The recognition rates (%) of various methods on 15-Scenes database

Method	KSPM [17]	ScSPM [11]	LLC [18]	GLP [19]	Boureau *et al.* [20]	Boureau *et al.* [21]	**DL-COPAR**
Accuracy	81.40	80.28	79.24	83.20	83.3	84.9	**85.37**

Table 4. The recognition rates (%) of various methods on Caltech101 dataset. 15 and 30 images pre class are randomly selected for training, shown in the second and third row, respectively.

Method	KSPM [17]	ScSPM [11]	LLC [18]	GLP [19]	Boureau *et al.* [20]	Boureau *et al.* [21]	**DL-COPAR**
15 training	56.44	67.00	65.43	70.34	-	-	**75.10**
30 training	64.40	73.20	73.44	82.60	77.3	75.70	**83.26**

than that of other SPM methods, $21 \times 1024 = 21504$, as a result of that we discard the codes corresponding to the commonality.

We compare the proposed DL-COPAR with several recent methods which are based on SPM and sparse coding. The direct comparisons are shown in Table 3. KSPM [17], which represents the popular nonlinear kernel SPM, using spatial-pyramid histograms and Chi-square kernels, achieves 81.40% accuracy. Compared with KSPM, ScSPM [11], as the baseline method in our paper, employs sparse coding and max pooling and linear SVM for classification. It achieves 80.28% classification rate. LLC [18] considers locality relations for sparse coding, achieving 79.24%. GLP [19] focuses on a more sophisticated pooling method, which learns a weight for SIFT descriptors in pooling process by enhancing discrimination and incorporates local spatial relations. With high computational cost (eigenvalue decomposition of large size matrix and gradient ascent process), GLP achieves 83.20% accuracy on this dataset. In [21], 84.9% classification accuracy is achieved through macrofeatures, which jointly encodes a small neighborhood SIFT descriptors and learns a discriminative dictionary.

Obviously, 85.37% accuracy achieved by our method is higher than that of all the other methods, and is also competitive with the current state-of-the art of 88.1% reported in [24] and 89.75% in [25]. It is worth noting that our DL-COPAR uses only the SIFT descriptors and solves the classical LASSO problem for sparse coding, while the method in [24] combines 14 different low-level features, and that in [25] involves a more complicated coding process called Laplacian sparse coding, which requires a well-estimated similarity matrix and multiple optimization stages.

4.6 Object Recognition

Caltech101 dataset [26] contains 9144 images in 101 object classes including animals, vehicles, flowers, etc, and one background category. Each class contains 31 to 800 images with significant variance in shape. As suggested by the original dataset and also by the previous studies, we randomly select 15 and 30 images respectively per class for training and report the classification accuracies averaged over the 102 classes. The size of the overall dictionary is set $K = 2048$ which is a popular set in the community, and DL-COPAR learns 17 visual words for each category and 314 for the shared pattern

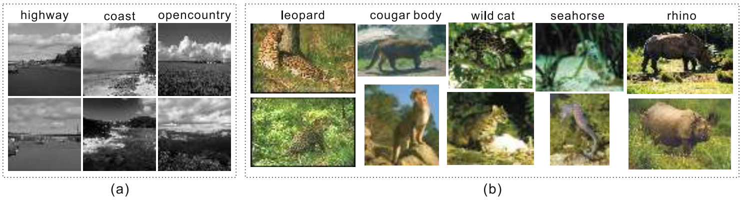

Fig. 7. Some example images from Scene15 and Caltech101. Obviously, despite the class-specific features, the images of different categories share some common patterns, such as the clouds in (a) and the flora background in (b).

pool, $17 \times 102 + 314 = 2048$ in total. Similar to scene recognition experiment, the final image representation vector ($21 \times 17 \times 102 = 36414$) is still much less than that of other SPM methods ($21 \times 2048 = 43008$), as a result of discarding the sparse codes corresponding to the commonality.

Table 4 demonstrates the direct comparisons. It is easy to see our DL-COPAR consistently outperforms other SPM-based methods, achieving 83.26% accuracy which is competitive to the state-of-the-art performance of 84.3% achieved by a group-sensitive multiple kernel method [27]. Compared with ScSPM, which is the baseline method in this experiment, DL-COPAR improves the classification performance with a margin of more than 8% and 10% when 15 and 30 images are randomly selected for training, respectively. As demonstrated by Fig. 7, we conjecture the reason why our DL-COPAR achieves such a high classification accuracy on both 15-scene and Caltech101 is that, for example, the complex background are intensively encoded over the learned commonality, whereas the part of the coefficient corresponding to the commonality is discarded. Thus, the omission of the common patterns will improve the classification performance.

5 Conclusion

Under the empirical observation that images (objects) from different categories usually share some common patterns which are not helpful for classification but essential for representation, we propose a novel approach based on dictionary learning to explicitly learn the shared pattern pool (the commonality) and the class-specific dictionaries (the particularity), dubbed **DL-COPAR**. Therefore, the combination of the particularity (corresponding to the specific class) and the commonality can faithfully represent the samples from this class, and the particularities are more discriminative and more compact for classification. Through experiments on a synthetic dataset and several public benchmarks for various applications, we can see the proposed DL-COPAR achieves very promising performances on these datasets.

References

1. Elad, M., Aharon, M.: Image denoising via learned dictionaries and sparse representation. In: CVPR (2006)
2. Bradley, D.M., Bagnell, J.A.: Differentiable sparse coding. In: NIPS (2008)

3. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Supervised dictionary learning. In: NIPS (2008)
4. Ramirez, I., Sprechmann, P., Sapiro, G.: Classification and clustering via dictionary learning with structured incoherence and shared features. In: CVPR (2010)
5. Zhang, Q., Li, B.: Discriminative k-svd for dictionary learning in face recognition. In: CVPR (2010)
6. Yang, M., Zhang, L., Feng, X., Zhang, D.: Fisher discrimination dictionary learning for sparse representation. In: ICCV (2011)
7. Wang, F., Lee, N., Sun, J., Hu, J., Ebadollahi, S.: Automatic group sparse coding. In: AAAI (2011)
8. Yang, M., Zhang, L., Yang, J., Zhang, D.: Metaface learning for sparse representation based face recognition. In: ICIP (2010)
9. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: NIPS (2007)
10. Aharon, M., Elad, M., Bruckstein, A.M.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Processing 54, 4311–4322 (2006)
11. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR (2009)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
13. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. PAMI (2001)
14. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. PAMI (2009)
15. Jiang, Z., Lin, Z., Davis, L.S.: Learning a discriminative dictionary for sparse coding via label consistent k-svd. In: CVPR (2011)
16. Huang, K., Aviyente, S.: Sparse representation for signal classification. In: NIPS (2006)
17. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognition natural scene categories. In: CVPR (2006)
18. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Learning locality-constrained linear coding for image classification. In: CVPR (2010)
19. Feng, J., Ni, B., Tian, Q., Yan, S.: Geometric ℓ_p norm feature pooling for image classification. In: CVPR (2011)
20. Boureau, Y.L., Roux, N.L., Bach, F., Ponce, J., LeCun, Y.: Ask the locals: multi-way local pooling for image recognition. In: ICCV (2011)
21. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR (2010)
22. Oliva, A., Torraba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelop. IJCV (2001)
23. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2005)
24. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large scale scene recognition from abbey to zoo. In: CVPR (2010)
25. Gao, S., Tsang, I.W.H., Chia, L.T., Zhao, P.: local features are not lonely - laplacian sparse coding for image classification. In: CVPR (2010)
26. Li, F.F., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: IEEE CVPR Workshop on Generative-Model Based Vision (2004)
27. Yang, J., Li, Y., Tian, Y., Duan, L., Gao, W.: Group-sensitive multiple kernel learning for object categorization. In: ICCV (2009)