Image Classification With Tailored Fine-Grained Dictionaries

Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Zechao Li, Yu-Gang Jiang, and Shuicheng Yan

Abstract— **In this paper, we propose a novel fine-grained dictionary learning method for image classification. To learn a high-quality discriminative dictionary, three types of multispecific subdictionaries, i.e., class-specific dictionaries (CSDs), universal dictionary (UD), and family-specific dictionaries (FSDs), are simultaneously uncovered. Here, CSDs and UD, respectively, model the patterns for each class and the patterns irrespective of any class. FSDs can help reveal the shared patterns between multiple image classes, by filling the gap between the patterns in CSDs and UD. The dependence among image classes is revealed by the shared FSDs, and a common FSD can be assigned to several classes to represent their residual. Finally, the most discriminative FSD for each class is identified by minimizing the sparse reconstruction error. Extensive experiments are conducted on different widely used data sets for image classification. The results demonstrate the superior performance of the proposed method over some state-of-the-art methods.**

Index Terms— **Class-specific dictionaries (CSDs), dictionary learning, family-specific dictionaries (FSDs), image classification, universal dictionary (UD).**

I. Introduction

SPARSE representation is inspired from the observation that an input image $\mathbf{y}$ can often be well approximated by a linear combination of a few representative samples from a type of redundant base, namely, $\mathbf{y} = \mathbf{Dx}$, where $\mathbf{D}$ is a dictionary evolved from the codebook [1] or vocabulary [2] and $\mathbf{x}$ is the sparse coefficient vector. Dictionary learning, as a particular sparse representation model, has been widely used for image/video coding [3], [4], image compression [5], image reconstruction [6], [7], and image classification [8]–[22]. It aims to learn a unified and compact dictionary from images to well reconstruct them. Traditionally, unsupervised dictionary learning adopts a given set of image samples to pursue the minimum error of sparse reconstruction [6], [23]–[27]. However, recent works have shown that better dictionaries can be found by supervised dictionary learning methods [10], [11], [18]. They explore the supervised information to learn discriminative dictionaries by bringing in the class-discriminative information [1], [8], [9], [12], [19], [21], [28]–[34]. Such supervised dictionary learning methods learn either an overall dictionary independent of any classes, or several subdictionaries related to different classes.

Empirically, images from different classes usually share some common patterns (nonclass-specific patterns) besides the class-specific patterns. To this end, [35] and [36] attempted to learn an additional universal dictionary (UD) to represent the common patterns appearing in some classes, e.g., the background in object recognition [35], [36] and handwritten digit recognition [21]. These patterns are referred to universal patterns in this paper.

These methods roughly assume that all the classes have the common nonclass-specific patterns, where these patterns contribute equivalently to the UD. However, this assumption underestimates the complexity of these patterns across different classes. Thus, they cannot well capture the subtle differences and shared patterns among the fine-grained visual classes while facing with fine-grained image classification tasks [37]. Fine-grained image classification refers to the task of classifying objects that belong to the same basic-level class (e.g., different ower species) and share similar shapes or visual appearances.

Recently, Gao *et al.* [38] attempted to learn a class-specific dictionary (CSD) for each class and a UD for all the classes in the fine-grained image classification task. However, some classes, rather than all classes, may have more similar shapes or visual appearances, which are difficult to be simply described by a CSD or a UD. For example, in the case of recognizing human and object interactions as shown in Fig. 1, eight example images sampled from the widely used People-Playing-Music-Instruments (PPMI) data set [39] have not only the universal patterns, e.g., people, but also some specific patterns partially shared by several classes rather than not all classes, such as some patterns of "guitar" existing in class "play guitar" and "with guitar." The patterns of "guitar" do not contribute the discriminative features to classifying "play guitar" and "with guitar," but contribute to distinguishing these

Manuscript received May 31, 2015; revised February 19, 2016 and July 18, 2016; accepted August 27, 2016. Date of publication September 8, 2016; date of current version February 13, 2018. This work was supported in part by the 973 Program of China under Project 2014CB347600, in part by the National Natural Science Foundation of China under Grant 61522203 and Grant 61402228, and in part by the National Ten Thousand Talent Program of China (Young Top-Notch Talent). This paper was recommended by Associate Editor P. Schelkens. *(Corresponding author: Xiangbo Shu.)*

X. Shu, J. Tang, and Z. Li are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: shuxb@njust.edu.cn; jinhuitang@njust.edu.cn; zechao.li@njust.edu.cn).

G.-J. Qi is with the Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: guo-jun.qi@ucf.edu).

Y.-G. Jiang is with the School of Computer Science, Fudan University, Shanghai, China (e-mail: ygj@fudan.edu.cn).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: eleyans@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2607345

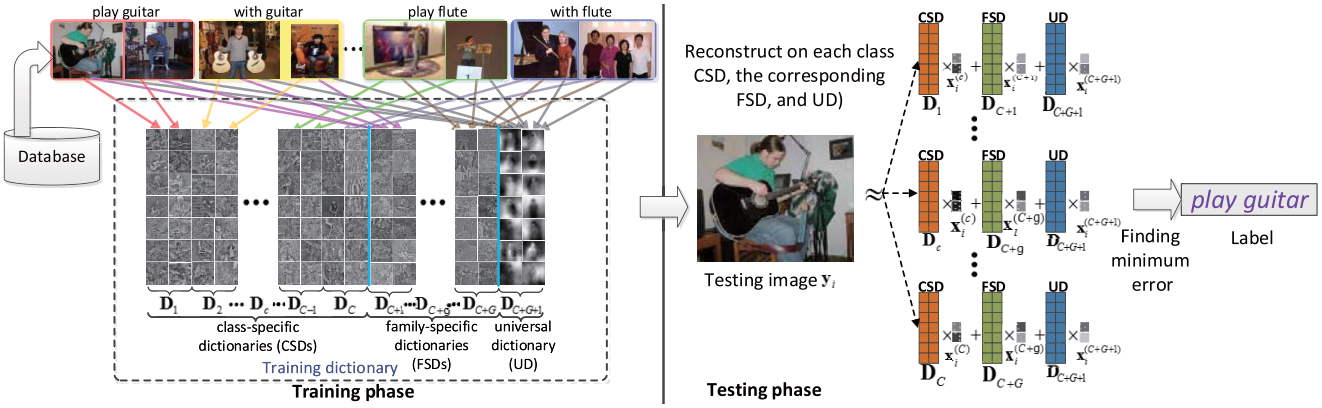


Fig. 1. Flowchart of the proposed FDL. $\mathbf{D}_1, \ldots, \mathbf{D}_C$ denote CSDs, $\mathbf{D}_{C+1}, \ldots, \mathbf{D}_{C+G}$ denote FSDs, and $\mathbf{D}_{C+G+1}$ denotes a UD. Given a testing image $\mathbf{y}_i$, we reconstruct it on each CSD, the corresponding FSD, and the UD, and then we obtain its label by finding the minimum reconstruction error. Note that: 1) all classes have the corresponding CSDs and share the UD $\mathbf{D}_{C+G+1}$ to encode their common background patterns, i.e., the indoor scene and people and 2) we also learn FSDs for some shared patterns among several classes. For example, the class "play guitar" and "with guitar" contribute to an FSD $\mathbf{D}_{C+1}$ for some patterns of "guitar," while the class "play flute" and "with flute" contribute to an FSD $\mathbf{D}_{C+G}$ for some patterns of "flute."

two classes from the other classes. These kinds of patterns (e.g., the patterns of "guitar" in the example), which are class-specific for basic-level classification but nonclass-specific for fine-grained classification, are called family-specific patterns that are distinguished from class-specific patterns and universal patterns. Therefore, it is necessary to discover the corresponding family-specific dictionaries (FSDs) to represent the family-specific patterns.

Toward this end, we propose to learn the hybrid dictionaries, including the FSDs, the CSDs for each class, and a UD. To the best of our knowledge, it is the first time to build the FSDs to reveal fine-grained dependence among classes. Fig. 1 illustrates the hybrid structure of the desired dictionaries. The eight example images from four classes have a similar indoor scene in the background, which is encoded in the UD. Although "guitar" is present in class "play guitar" and "with guitar," there is no "guitar" in class "play flute" and "with flute." Similarly, "flute" appears in class "play flute" and "with flute," while it is absent in class "play guitar" and "with guitar." Clearly, the patterns of "guitar" and "flute" are neither the class-specific patterns nor the universal patterns for the fine-grained image classification. Instead, they are the family-specific patterns.

In this paper, we propose a fine-grained dictionary learning (FDL) method to learn a high-quality discriminative dictionary by jointly learning the CSDs, FSDs, and UD. The proposed FDL learns the overall dictionary in a supervised way and explicitly learns the fine-grained dictionaries (CSDs, FSDs, and UD). The sparse reconstruction cost is minimized for the separate family-specific, class-specific, and universal patterns of training images to iteratively group all the classes into several families. To verify the effectiveness of the proposed method, extensive experiments are conducted on several widely used data sets. The results demonstrate its superior performance compared with other state-of-the-art dictionary learning methods for both the basic-level image classification tasks and the fine-grained image classification tasks.

The main contributions of this paper can be summarized as follows.

1) We propose to learn the fine-grained dictionaries by leveraging the CSDs, FSDs, and UD simultaneously. It is the first time to explicitly discover the FSDs to reveal the fine-grained dependence among classes.
2) We propose to group all classes into several families iteratively by minimizing the sparse reconstruction cost for the separate family-specific, class-specific, and universal patterns of the training images. An efficient algorithm is developed to solve the optimization problem.
3) Existing dictionary learning methods can be seen as special cases of the proposed method. Compared with previous works, the proposed method is more suitable for real applications and improves the performance on various kinds of classification tasks, especially for fine-grained image classification.

The rest of this paper is organized as follows. Section II introduces the preliminary works. We present the proposed FDL method in Section III, followed by the optimization algorithm in Section IV. The experimental results and analysis are presented in Section V. We give the analysis of time complexity in Section VI. Finally, the conclusion and future work are presented in Section VII.

II. Preliminaries: Sparse Representation and Dictionary Learning-Based Classification

Throughout this paper, the scalars, vectors, and matrices are denoted by lowercase letters, bold lowercase characters, and bold uppercase characters, respectively. When $\mathbf{A}$ is a matrix in $\mathbb{R}^{d\times K}$, $\mathbf{A}_c = [\mathbf{a}_1^c, \ldots, \mathbf{a}_{K_c}^c] \in \mathbb{R}^{d\times K_c}$ and $\mathbf{A}^{(c)} = [\mathbf{a}_1^{(c)}; \ldots; \mathbf{a}_{d_c}^{(c)}] \in \mathbb{R}^{d_c\times K}$ denote a column-block matrix and a row-block matrix of $\mathbf{A}$, respectively, where $\mathbf{a}_1^c \in \mathbb{R}^{d\times 1}$ represents a column vector of and $\mathbf{a}_1^{(c)} \in \mathbb{R}^{1\times K}$ is the row-block vector of $\mathbf{a} \in \mathbb{R}^K$. $\mathbf{A}_c^{(c)} \in \mathbb{R}^{d_c\times K_c}$ is a column- and row-block matrix of $\mathbf{A}$. Besides, we denote $\mathbf{I_m}$ by the identity matrix of size $m \times m$. $\mathbf{0}_{m\times n}$ is the zero matrix in $\mathbb{R}^{m\times n}$, and $\mathbf{A^T}$ is the transposed matrix of $\mathbf{A}$.

The principle of sparse representation [40], [41] is that a given image can be collaboratively encoded over

a codebook/dictionary with some sparsity constraints, such as ℓ_1-norm and ℓ_0-norm minimization. Sparse-representation-based classification (SRC) applies the sparse coding to the classification task [42]. Suppose there are C classes of data. Let $\mathbf{Y}_c \in \mathbb{R}^{d\times N_c}$ denote the training samples of the cth class. Then we can define the dictionary $\mathbf{A} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_c, \ldots, \mathbf{Y}_C] \in \mathbb{R}^{d\times N}$, where $N = \sum_{i=1}^{C} N_c$. Denote $\mathbf{y}$ by an input image. The SRC model is formulated as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{1}$$

where λ is a regularization parameter to balance the reconstruction error and the sparsity. The classification strategy of SRC is

$$c = \arg\min_{c} \|\mathbf{y} - \mathbf{Y}_c\hat{\mathbf{x}}^{(c)}\|_2^2. \tag{2}$$

In general, SRC directly uses the original training images as the dictionary, which will lead to the redundance of the dictionary, and even inaccurate results due to the noisy and trivial information. Empirically, the size of the predefined dictionary in SRC is usually large, which increases the computational complexity and memory. Therefore, the class-oriented dictionary learning method is proposed in [21] and [18]. It assumes that the learned discriminative dictionary should be composed of all the CSDs. We denote that $\mathbf{Y}_c \in \mathbb{R}^{d\times N_c}$ and $\mathbf{X}_c \in \mathbb{R}^{K_c\times N_c}$ are the collection of training samples of class c and the corresponding coefficients, respectively, where K_c is the number of atoms of dictionary. The CSD learning for the cth class is expressed as

$$\begin{aligned} &\min_{\mathbf{D}_c,\mathbf{X}_c} \|\mathbf{Y}_c - \mathbf{D}_c\mathbf{X}_c\|_F^2 + \lambda\|\mathbf{X}_c\|_1 \\ &\text{s.t. } \left\|\mathbf{d}_l^c\right\|_2 \le 1 \quad \forall l = 1, 2, \ldots, K_c \end{aligned} \tag{3}$$

where $\mathbf{D}_c = [\mathbf{d}_1^c, \ldots, \mathbf{d}_{K_c}^c] \in \mathbb{R}^{d\times K_c}$, $\|\mathbf{X}_c\|_1 = \sum_{i=1}^{N_c} \|\mathbf{x}_i^c\|_1$, and $\mathbf{d}_l^{(c)} \in \mathbb{R}^{d\times 1}$ is the lth column (atom) of the cth CSD $\mathbf{D}_c$. The constraint on the ℓ_2-norm of all the atoms $\mathbf{d}_l^{(c)}$ prevents the optimal value of $\mathbf{D}_c$ from being arbitrarily large, which leads to small sparse coefficients $\mathbf{X}_c$.[1] The desired $\mathbf{D}_c$ ($c = 1, 2, \ldots, C$) corresponding to a specific class should be incoherent with other classes. In order to increase this incoherence among the CSDs, Ramirez *et al.* [21] forced the incoherence penalty between pairwise CSDs into the optimization problem (3) via the inner products between each pair of atoms from different dictionaries

$$\begin{aligned} \min_{\mathbf{D}_c,\mathbf{X}_c} &\sum_{c=1}^{C} \left\{\|\mathbf{Y}_c - \mathbf{D}_c\mathbf{X}_c\|_F^2 + \lambda\|\mathbf{X}_c\|_1\right\} \\ &+\beta\sum_{c=1}^{C}\sum_{j=1,j\neq c}^{C} \left\|\mathbf{D}_c^T\mathbf{D}_j\right\|_F^2 \end{aligned} \tag{4}$$

where β is a nonnegative parameter to control the incoherence penalty. The last term forces the coherence of different CSDs to be small. All the learned CSDs are concatenated into one overall dictionary $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_C]$.

[1]For convenience, we omit the constraint on the ℓ_2-norm of each atom, but it always exists for all the objective functions about dictionary learning throughout this paper.

III. Fine-Grained Dictionary Learning

In this section, an objective function is formulated to learn the tailored fine-grained dictionaries in allusion to the fact that one image contains the class-specific, family-specific, and universal patterns.

A. Unified Dictionary With Sparse Reconstruction

Suppose we have N training images from C classes, denoted by $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N] \in \mathbb{R}^{d\times N}$, where $\mathbf{y}_i$ is the dD feature descriptor of the ith image of the training data. One fundamental principle of our dictionary learning model is to guarantee that the learned unified dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_k, \ldots, \mathbf{d}_K] \in \mathbb{R}^{d\times K}$ can represent each image $\mathbf{y}_i$ well by sparse weighted linear combinations of atoms

$$\min_{\mathbf{D},\mathbf{X}} \sum_{i=1}^{N} \left\{\|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda\|\mathbf{x}_i\|_1\right\}. \tag{5}$$

This optimization problem is the classical dictionary learning model, which can guarantee the representation power. However, the learned dictionary $\mathbf{D}$ does not have any discriminative information.

B. Tailored Dictionaries With Specific Orientation

Although the basic dictionary learning model in (5) can be directly applied to the classification task, some works [10], [13], [43] confirm that better results can be obtained when the dictionary is tuned according to some certain tasks. In this paper, our goal is to design a high-quality discriminative dictionary learning model, which can learn a pure class-discriminative dictionary by exploring the general class information. Based on the fact that one image explicitly contains three types of subpatterns, i.e., class-specific patterns, family-specific patterns, and universal patterns, we can represent the overall dictionary $\mathbf{D}$ as

$$\begin{aligned} \mathbf{D} = [&\mathbf{D}_1, \ldots, \mathbf{D}_c, \ldots, \mathbf{D}_C, \mathbf{D}_{C+1}, \ldots, \mathbf{D}_{C+g}, \ldots, \\ &\mathbf{D}_{C+G}, \mathbf{D}_{C+G+1}] \end{aligned} \tag{6}$$

where $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_C$ denote the CSDs, $\mathbf{D}_{C+1}, \ldots, \mathbf{D}_{C+g}, \ldots, \mathbf{D}_{C+G}$ denote the FSDs, and $\mathbf{D}_{C+G+1}$ denotes the UD. $\mathbf{D}_i \in \mathbb{R}^{d\times K_i}$ ($i = 1, \ldots, C + G + 1$), $K = \sum_{i=1}^{C+G+1} K_i$, and $g \in \{1, \ldots, G\}$ is a family cluster index. Specifically, we assume that the images of the cth class $\mathbf{Y}_c = [\mathbf{y}_1^c, \ldots, \mathbf{y}_i^c, \ldots, \mathbf{y}_{N_c}^c]$ can be reconstructed by three types of subdictionaries

$$\mathbf{Y}_c \approx \mathbf{D}_c\mathbf{X}_c^{(c)} + \mathbf{D}_{C+g}\mathbf{X}_c^{(C+g)} + \mathbf{D}_{C+G+1}\mathbf{X}_c^{(C+G+1)}. \tag{7}$$

As shown in Fig. 2, the image $\mathbf{y}_i^c$ from the class "play guitar" in the PPAMI data set belongs to the cth class, and the cth class is grouped into the gth family cluster (a total of G family clusters). The image $\mathbf{y}_i^c$ contains three types of subpatterns, i.e., class-specific patterns (action of playing guitar), universal patterns (indoor scene, as the background), as well as the family-specific patterns (guitar, as the residual).

The correspondences between class-specific patterns and the CSD are known, namely, the data class prior. And the universal patterns of all classes also correspond to a UD. However, the

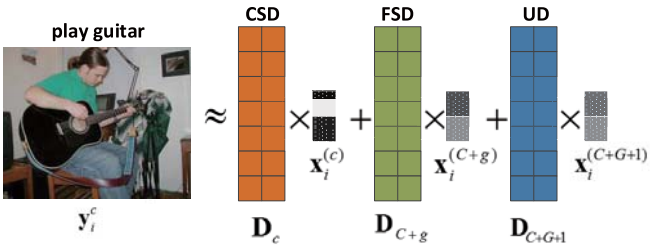


Fig. 2. Example of the motivation. An image from the cth class ("play guitar") and the gth family cluster should be approximated by linearly combining three types of multispecific subdictionaries, i.e., CSD, FSD, and UD.

explicit correspondences between family-specific patterns and the FSD are unavailable, since we cannot obtain any prior relationships between the training image classes and the FSDs. A clustering framework [21] was proposed based on dictionary learning instead of searching the centroid of the set that well fits the data, such as nonnegative spectral clustering [44], [45]. Therefore, we consider to use an unsupervised clustering model to find the optimal family cluster for each class by minimizing the reconstruction error. As in a general clustering algorithm, we formally define the number of clusters as G and obtain a set of G FSDs $\{\mathbf{D}_{C+1}, \ldots, \mathbf{D}_{C+g}, \ldots, \mathbf{D}_{C+G}\}$ by the optimization problem

$$\min_{\mathbf{D},\mathbf{X}} \sum_{c=1}^{C} \left\{ \min_{g} \left\| \mathbf{Y}_c - \mathbf{D}_c \mathbf{X}_c^{(c)} - \mathbf{D}_{C+g} \mathbf{X}_c^{(C+g)} - \mathbf{D}_{C+G+1} \mathbf{X}_c^{(C+G+1)} \right\|_F^2 + \lambda \|\mathbf{X}_c\|_1 \right\} \quad (8)$$

where $g = 1, \ldots, G$.

C. Objective Function

Formula (5) mainly focuses on how to learn an overall dictionary that fits all training images as well as possible, while formula (8) pursues a sparse representation for each class through three subdictionaries. In order to increase the structure incoherence among all the subdictionaries, we also introduce the incoherence penalty between each pair of CSDs [21]. Therefore, to derive a high-quality discriminative dictionary, the objective function of the proposed FDL method is formulated by combining (5) and (8) and the incoherence penalty

$$\begin{aligned} f &= \min_{\mathbf{D},\mathbf{X}} \sum_{i=1}^{N} \left\{ \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_F^2 + \lambda \|\mathbf{x}_i\|_1 \right\} + \sum_{c}^{C} f_c(\mathbf{D}_{C+g}) \\ &\quad + \beta \sum_{i=1}^{C+G+1} \sum_{j=1, j\neq i}^{C+G+1} \left\| \mathbf{D}_i^T \mathbf{D}_j \right\|_F^2 \\ &= \min_{\mathbf{D},\mathbf{X}} \sum_{c}^{C} \left\{ \|\mathbf{Y}_c - \mathbf{D}\mathbf{X}_c\|_F^2 + f_c(\mathbf{D}_{C+g}) + \lambda \|\mathbf{X}_c\|_1 \right\} \\ &\quad + \beta \sum_{i=1}^{C+G+1} \sum_{j=1, j\neq i}^{C+G+1} \left\| \mathbf{D}_i^T \mathbf{D}_j \right\|_F^2 \end{aligned} \quad (9)$$

where β and λ are the parameters that control the sparsity penalty and the incoherence penalty, respectively. In (9), we omit the repeating sparsity penalty terms, and transform (8) into an unsupervised learning term $f_c(\mathbf{D}_{C+g})$, namely, for $\forall c = 1, \ldots, C$

$$f_c(\mathbf{D}_{C+g}) = \min_{g} \left\| \mathbf{Y}_c - \mathbf{D}_c \mathbf{X}_c^{(c)} - \mathbf{D}_{C+g} \mathbf{X}_c^{(C+g)} - \mathbf{D}_{C+G+1} \mathbf{X}_c^{(C+G+1)} \right\|_F^2. \quad (10)$$

However, even if we force $\mathbf{D}_c$, $\mathbf{D}_{C+g}$ and $\mathbf{D}_{C+G+1}$ to well approximate the data from the cth class and the gth cluster, other part of the products of subdictionaries and the sparse coefficient may still be nonzero. Yang *et al.* [18] and Kong and Wang [35] forced the products of other subdictionaries (except for the ones related to the current class) and corresponding subcoefficients to be nearly zero. In our work, if the cth class is grouped into the gth cluster, we employ $\sum_{i,i\neq c,i\neq g,i\neq C+G+1}^{C+G+1} \|\mathbf{X}_c^{(i)}\|_F^2$ to make some unrelated parts of the coefficients be zero.

For $\forall i = 1, \ldots, C+G+1$, denote $\mathbf{H}_i = [\mathbf{h}_1^{(i)}; \ldots; \mathbf{h}_i^{(i)}; \ldots; \mathbf{h}_{K_i}^{(i)}] \in \mathbb{R}^{K_i \times K}$ and $\mathbf{H}_{/i} = [\mathbf{H}_1; \ldots; \mathbf{H}_{i-1}; \mathbf{0}_{K_i \times K}; \mathbf{H}_{i+1}; \ldots; \mathbf{H}_{C+G+1}] \in \mathbb{R}^{K \times K}$, where $\mathbf{h}_j^{(i)} \in \mathbb{R}^{1 \times K}$ is a row vector

$$\mathbf{h}_j^{(i)} = \left[\underbrace{0, \ldots, 0}_{\sum_{m=1}^{i-1} K_m}, \underbrace{\overbrace{0, \ldots, 0}^{j-1}, \overset{j}{\overset{\downarrow}{1}}, \overbrace{0, \ldots, 0}^{K_i - j}}_{K_i}, \underbrace{0, \ldots, 0}_{\sum_{m=i+1}^{C+G+1-c} K_m} \right]. \quad (11)$$

Then for $\forall c = 1, \ldots, C$, we conduct $\mathbf{X}_c^{(c)} = \mathbf{H}_c \mathbf{X}_c$ and $\mathbf{D}_c = \mathbf{D}\mathbf{H}_c^T$ through some matrix computation steps. For mathematical brevity, let

$$\sum_{i,i\neq c,i\neq g,i\neq C+G+1}^{C+G+1} \left\| \mathbf{X}_c^{(i)} \right\|_F^2 = \|\mathbf{H}_{/(c,C+g,C+G+1)} \mathbf{X}_c\|_F^2 \quad (12)$$

where $\mathbf{H}_{/(c,C+g,C+G+1)} = [\mathbf{H}_1; \ldots; \mathbf{H}_{c-1}; \mathbf{0}_{K_c \times K}; \mathbf{H}_{c+1}; \ldots; \mathbf{H}_C; \mathbf{H}_{C+1}; \ldots; \mathbf{H}_{C+g-1}; \mathbf{0}_{K_{C+g} \times K}; \mathbf{H}_{C+g+1} \ldots; \mathbf{H}_{C+G}; \mathbf{0}_{K_{C+G+1} \times K}]$. We add the constraint of (12) and redefine (10): for $c = 1, \ldots, C$

$$\begin{aligned} &f_c(\mathbf{D}_{C+g}, \mathbf{H}_{/(c,C+g,C+G+1)}) \\ &= \min_{g} \left\{ \delta \|\mathbf{H}_{/(c,C+g,C+G+1)} \mathbf{X}_c\|_F^2 \right. \\ &\quad \left. + \left\| \mathbf{Y}_c - \mathbf{D}_c \mathbf{X}_c^{(c)} - \mathbf{D}_{C+g} \mathbf{X}_c^{(C+g)} - \mathbf{D}_{C+G+1} \mathbf{X}_c^{(C+G+1)} \right\|_F^2 \right\} \end{aligned} \quad (13)$$

where the parameter δ balances the reconstruction cost and coefficients penalty. To select the optimal g, the first term in (13) is to explore minimum of subcoefficients $\{\mathbf{X}_{C+1}, \ldots, \mathbf{X}_{C+g-1}, \mathbf{X}_{C+g+1}, \ldots, \mathbf{X}_{C+G}\}$, and the second term in (13) measures the reconstitution cost. Thus, we renew the objective function f in (9) to obtain the final objective function J of the proposed FDL method

$$\begin{aligned} J &= \min_{\mathbf{D},\mathbf{X}} \sum_{c}^{C} \left\{ \|\mathbf{Y}_c - \mathbf{D}\mathbf{X}_c\|_F^2 + f_c(\mathbf{D}_{C+g}, \mathbf{H}_{/(c,C+g,C+G+1)}) \right. \\ &\quad \left. + \lambda \varphi(\mathbf{X}_c) \right\} + \beta \sum_{i=1}^{C+G+1} \sum_{j=1, j\neq i}^{C+G+1} \left\| \mathbf{D}_i^T \mathbf{D}_j \right\|_F^2. \end{aligned} \quad (14)$$

The overall dictionary learned via the above process has the following properties.

1) Since the learned dictionary is adaptive to the specifics of each class, it can lead to a better representation of each image with strict sparsity.

Algorithm 1 FDL

Input: C class training data $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_c, \ldots, \mathbf{Y}_C] \in \mathbb{R}^{d\times N}$, the size K_i $(i = 1, \ldots, C+G+1)$ of all subdictionaries, parameters δ, β, λ, and G.
Output: dictionary $\mathbf{D}$ and sparse coefficient $\mathbf{X}$.
Ensure: $\|\mathbf{d_i}\|_2 = 1$[1], for $\forall i = 1, \ldots, K$.
Initialize: $\mathbf{D}_i$, $i = 1, \ldots, C+G+1$.
repeat
  updating the coefficient matrix $\mathbf{X}$ by solving (15);
  finding the clusters $\hat{g}$ of c $(c = 1, \ldots, C)$ with (17);
  updating the atoms of $\mathbf{D}_c$ by fixing $\mathbf{X}$ and $\mathbf{D}_{/c}$ with (21), for $c = 1, \ldots, C$;
  updating the atoms of $\mathbf{D}_{C+g}$ by fixing $\mathbf{X}$ and $\mathbf{D}_{/C+g}$ with (24), for $g = 1, \ldots, G$;
  updating the atoms of $\mathbf{D}_{C+G+1}$ by fixing $\mathbf{X}$ and $\mathbf{D}_{/C+G+1}$ with (26).
until Convergence

[1]When updating each atom $\mathbf{d}_i$, we normalize it to satisfy this constraint, i.e., $\mathbf{d}_l = \mathbf{d}_l/\|\mathbf{d_l}\|_\mathbf{2}$, and the corresponding coefficient is multiplied with $\|\mathbf{d}_l\|_2$, i.e., $\mathbf{X}^{(l)} = \|\mathbf{d}_l\|_2\mathbf{X}^{(l)}$.

2) FSDs and UD can capture the nonclass-specific patterns, which leads to that the class-specific patterns are purely revealed by the CSDs. In other words, FSDs and UD make CSDs be more compact and class discriminative.
3) The obtained sparse coefficients are discriminative, which can improve the discriminative ability of the dictionary. Thus, the dictionaries and the sparse coefficients can boost the class-discriminative ability of each other.

IV. Optimization Procedure

The objective function J in (14) is convex with respect to $\mathbf{D}$ and $\mathbf{X}$ separately. The optimization of J can be iteratively solved through three alternative subprocedures of optimization. Specifically, we fix other variables when updating one variable. The process iterates until convergence. The convergence criterion is that the iteration steps shall end when the relative change of all variables is smaller than a predefined threshold. Algorithm 1 is summarized the algorithm details.

A. Updating $\mathbf{X}$

When updating $\mathbf{X}$, we fix $\mathbf{D}$ and select the terms related to $\mathbf{X}$ in the objective function J

$$\begin{aligned} &J(\mathbf{X}) \\ &= \min_{\mathbf{X}} \sum_{c}^{C} \left\{ \begin{array}{l} \left\| \mathbf{Y}_c - \mathbf{D}_c\mathbf{X}_c^{(c)} - \mathbf{D}_{\widehat{C+g}}\mathbf{X}_c^{(\widehat{C+g})} \right. \\ \left. - \mathbf{D}_{C+G+1}\mathbf{X}_c^{(C+G+1)} \right\|_F^2 + \|\mathbf{Y}_c - \mathbf{D}\mathbf{X}_c\|_F^2 \\ + \delta\|\mathbf{H}_{/(c,\widehat{C+g},C+G+1)}\mathbf{X}_c\|_F^2 + \lambda\|\mathbf{X}_c\|_1 \end{array} \right\} \\ &= \min_{\mathbf{X}} \sum_{c}^{C} \left\{ \begin{array}{l} \left\| \begin{bmatrix} \mathbf{Y}_c \\ \mathbf{Y}_c \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{D}\hat{\Gamma}_c \\ \mathbf{D} \\ \sqrt{\delta}\mathbf{H}_{/(c,\widehat{C+g},C+G+1)} \end{bmatrix} \mathbf{X}_c \right\|_F^2 \\ + \lambda\|\mathbf{X}_c\|_1 \end{array} \right\} \end{aligned} \tag{15}$$

where $\widehat{C+g}$ (with a hat notation) denotes that the corresponding $\mathbf{Y}_c$ in class c is grouped into the cluster g, $\mathbf{F}_c = \mathbf{H}_c^T\mathbf{H}_c$, $\hat{\Gamma}_c = \mathbf{F}_c + \mathbf{F}_{\widehat{C+g}} + \mathbf{F}_{C+G+1}$. To ensure the atoms of dictionaries to be ℓ_2-normalized, let $\tilde{\mathbf{Y}}_c = [\mathbf{Y}_c; \mathbf{Y}_c; \mathbf{0}]$, $\tilde{\mathbf{D}} = (1/\sqrt{2})[\mathbf{D}\hat{\Gamma}_c; \mathbf{D}; \sqrt{\delta}\mathbf{H}_{/(c,\widehat{C+g},C+G+1)}]$, $\tilde{\mathbf{X}}_c = \sqrt{2}\mathbf{X}_c$. Then, we obtain a simple form of the objective function $J(\mathbf{X})$

$$J(\tilde{\mathbf{X}}) = \min_{\mathbf{X}} \sum_{c}^{C} \left\{ \|\tilde{\mathbf{Y}}_c - \tilde{\mathbf{D}}\tilde{\mathbf{X}}_c\|_F^2 + \lambda\|\tilde{\mathbf{X}}_c\|_1 \right\}. \tag{16}$$

Note that $J(\tilde{\mathbf{X}})$ drops into a classical Lasso problem, which is effectively solved by the SPAMS toolbox.[2]

B. Clustering for g

When implementing the clustering to find the cluster g for each class, we fix $\mathbf{D}$ and $\mathbf{X}$, and omit the irrelevant terms in (14). For $\forall c = 1, \ldots, C$, its optimal cluster is

$$\begin{aligned} \hat{g} = \arg\min_{g} \Big\{ &\Big\| \mathbf{Y}_c - \mathbf{D}_c\mathbf{X}_c^{(c)} - \mathbf{D}_{C+g}\mathbf{X}_c^{(C+g)} \\ &- \mathbf{D}_{C+G+1}\mathbf{X}_c^{(C+G+1)} \Big\|_F^2 \\ &+ \delta\|\mathbf{H}_{/(c,C+g,C+G+1)}\mathbf{X}_c\|_F^2 \Big\}. \end{aligned} \tag{17}$$

After implementation of this clustering process, the $\hat{g}$ for class c $(c = 1, \ldots, C)$ is known in the current updating procedure.

C. Updating $\mathbf{D}$

We update $\mathbf{D}$ by fixing $\mathbf{X}$. Specifically, we update each subdictionary $\mathbf{D}_i$ $(i = 1, \ldots, C+G+1)$, respectively, while further fixing other subdictionaries. Similar to [35] and [21], we further propose to update $\mathbf{D}_i$ $(i = 1, \ldots, C+G+1)$ atom by atom. The detailed optimization procedure is divided into three ways.

D. Updating $\mathbf{D}_c$ $(c = 1, 2, \ldots, C)$

We omit the terms that are independent of $\mathbf{D}_c$ $(c = 1, 2, \cdots, C)$ from (14)

$$\begin{aligned} &J(\mathbf{D}_c) \\ &= \arg\min_{\mathbf{D}} \sum_{i}^{C} \left\{ \|\mathbf{Y}_i - \mathbf{D}\mathbf{X}_i\|_F^2 \right\} + 2\beta \sum_{\substack{j=1, \\ j\neq c}}^{C+G+1} \left\| \mathbf{D}_c^T\mathbf{D}_j \right\|_F^2 \\ &\quad + \left\| \mathbf{Y}_c - \mathbf{D}_c\mathbf{X}_c^{(c)} - \mathbf{D}_{\widehat{C+g}}\mathbf{X}_c^{(\widehat{C+g})} - \mathbf{D}_{C+G+1}\mathbf{X}_c^{(C+G+1)} \right\|_F^2. \end{aligned} \tag{18}$$

Denote $\mathbf{D}_c = [\mathbf{d}_1^{(c)}, \ldots, \mathbf{d}_l^{(c)}, \ldots, \mathbf{d}_{K_c}^{(c)}] \in \mathbb{R}^{d\times K_c}$. We update one atom $\mathbf{d}_l^{(c)}$ while fixing other atoms of $\mathbf{D}_c$. Hence, we get the objective function $J(\mathbf{d}_l^{(c)})$ of $\mathbf{d}_l^{(c)}$

$$\begin{aligned} J\left(\mathbf{d}_l^{(c)}\right) = \arg\min_{\mathbf{d}_l^{(c)}} \sum_{i=1}^{C} \Big\{ &\Big\| \mathbf{Y}_i - \mathbf{D}\mathbf{H}_{/c}^T\mathbf{H}_{/c}\mathbf{X}_i - \tilde{\mathbf{d}}_c\mathbf{H}_c\mathbf{X}_i \\ &- \mathbf{d}_l^{(c)}\mathbf{e}_l^c\mathbf{H}_c\mathbf{X}_i \Big\|_F^2 \Big\} \\ &+ \Big\| \mathbf{Y}_c - \tilde{\mathbf{d}}_c\mathbf{H}_c\mathbf{X}_c - \mathbf{d}_l^{(c)}\mathbf{e}_l^c\mathbf{H}_c\mathbf{X}_c \\ &- \mathbf{D}_{\widehat{C+g}}\mathbf{X}_c^{(\widehat{C+g})} - \mathbf{D}_{C+G+1}\mathbf{X}_c^{(C+G+1)} \Big\|_F^2 \\ &+ 2\beta \left\| \left(\mathbf{d}_l^{(c)}\right)^T \mathbf{D}\mathbf{H}_{/c}^T \right\|_F^2 \end{aligned} \tag{19}$$

where $\tilde{\mathbf{d}}_c = (\sum_{m=1,m\neq l}^{K_c} \mathbf{d}_m^{(c)}\mathbf{e}_m^c)$, and $\mathbf{e}_l^c$ is a row vector, of which the lth element is set to 1 and all other elements

[2]http://spams-devel.gforge.inria.fr/

are set to 0. Based on the notation definition of (29) in the Appendix, we rewrite (19)

$$J\left(\mathbf{d}_l^{(c)}\right) = \arg\min_{\mathbf{d}_l^{(c)}} \sum_{i=1}^{C} \left\{\left\|\mathbf{U}_i - \mathbf{d}_l^{(c)}\mathbf{r}_i\right\|_F^2\right\} + \left\|\mathbf{Z}_c - \mathbf{d}_l^{(c)}\mathbf{r}_c\right\|_F^2 + 2\beta\left\|\left(\mathbf{d}_l^{(c)}\right)^T\mathbf{V}_c\right\|_F^2. \quad (20)$$

Taking partial derivatives of $J(\mathbf{d}_l^{(c)})$ with respect to $\mathbf{d}_l^{(c)}$ and making them equal to zero, we obtain the updating rule as

$$\mathbf{d}_l^{(c)} = \left(\mathbf{A} + \mathbf{r}_c\mathbf{r}_c^T + 2\beta\mathbf{V}_c\mathbf{V}_c^T\right)^{-1}\left(\mathbf{B} + \mathbf{Z}_c\mathbf{r}_c^T\right) \quad (21)$$

where $\mathbf{A} = \sum_{i=1, i\neq c}^{C} \mathbf{r}_i\mathbf{r}_i^T$ and $\mathbf{B} = \sum_{i=1, i\neq c}^{C} \mathbf{U}_i\mathbf{r}_i^T$.

E. Updating $\mathbf{D}_C + g$ $(g = 1, 2, \ldots, G)$

When updating $\mathbf{D}_{C+g}$ $(g = 1, \ldots, G)$, we fix the other subdictionaries $\mathbf{D}_{/C+g}$. Similarly, let $\mathbf{D}_{C+g} = [\mathbf{d}_1^{(C+g)}, \ldots, \mathbf{d}_l^{(C+g)}, \ldots, \mathbf{d}_{K_{C+g}}^{(C+g)}] \in R^{d\times K_{C+g}}$, and we can update $\mathbf{d}_l^{(C+g)}$ while fixing the other atoms in $\mathbf{D}_{C+g}$. We also drop the unrelated terms of $\mathbf{D}_{C+g}$ in (14), and get the objective function J as

$$\begin{aligned} J\left(\mathbf{d}_l^{(C+g)}\right) &= \arg\min_{\mathbf{d}_l^{(C+g)}} \sum_{i=1}^{C} \left\{\left\|\mathbf{Y}_i - \mathbf{D}\mathbf{H}_{/C+g}^T\mathbf{H}_{/C+g}\mathbf{X}_i - \tilde{\mathbf{d}}_{C+g}\mathbf{H}_{/C+g}\mathbf{X}_i - \mathbf{d}_l^{(C+g)}\mathbf{e}_l^{C+g}\mathbf{H}_{/C+g}\mathbf{X}_i\right\|_F^2\right\} \\ &+ \sum_{i'\in\mathcal{S}_g} \left\{\left\|\mathbf{Y}_{i'} - \mathbf{D}_{i'}\mathbf{X}_{i'}^{(i')} - \mathbf{D}_{C+G+1}\mathbf{X}_{i'}^{(C+G+1)} + -\tilde{\mathbf{d}}_{C+g}\mathbf{H}_{/C+g}\mathbf{X}_{i'} - \mathbf{d}_l^{(C+g)}\mathbf{e}_l^{C+g}\mathbf{H}_{/C+g}\mathbf{X}_{i'}\right\|_F^2\right\} \\ &+ 2\beta\left\|\left(\mathbf{d}_l^{(C+g)}\right)^T\mathbf{D}\mathbf{H}_{/C+g}^T\right\|_F^2 \end{aligned} \quad (22)$$

where $\tilde{\mathbf{d}}_{C+g} = (\sum_{m=1, m\neq l}^{K_{C+g}} \mathbf{d}_m^{(C+g)}\mathbf{e}_m^{C+g})$ and $\mathbf{e}_l^c$ is a row vector, of which the lth element is set to 1 and all other elements are set to 0. Based on the definition of (30) in the Appendix, we rewrite (22) as

$$\begin{aligned} J\left(\mathbf{d}_l^{(C+g)}\right) &= \arg\min_{\mathbf{d}_l^{(C+g)}} \sum_{i=1}^{C} \left\{\left\|\mathbf{U}_i - \mathbf{d}_l^{(C+g)}\mathbf{r}_i\right\|_F^2\right\} \\ &+ 2\beta\left\|\left(\mathbf{d}_l^{(C+g)}\right)^T\mathbf{V}_{C+g}\right\|_F^2 + \sum_{i'\in\mathcal{S}_g} \left\{\left\|\mathbf{Z}_{i'} - \mathbf{d}_l^{(C+g)}\mathbf{r}_{i'}\right\|_F^2\right\}. \end{aligned} \quad (23)$$

Letting $\partial J(\mathbf{d}_l^{(C+g)})/\partial\mathbf{d}_l^{(C+g)} = 0$, we obtain the updating rule as

$$\mathbf{d}_l^{(C+g)} = \left(\mathbf{A} + \mathbf{Q} + 2\beta\mathbf{V}_{C+g}\mathbf{V}_{C+g}^T\right)^{-1}(\mathbf{B} + \mathbf{P}) \quad (24)$$

where $\mathbf{A} = \sum_{i=1}^{C} \mathbf{r}_i\mathbf{r}_i^T$, $\mathbf{B} = \sum_{i=1}^{C} \mathbf{U}_i\mathbf{r}_i^T$, $\mathbf{Q} = \sum_{i'\in\mathcal{S}_g} \mathbf{r}_{i'}\mathbf{r}_{i'}^T$, and $\mathbf{P} = \sum_{i'\in\mathcal{S}_g}^{C} \mathbf{Z}_{i'}\mathbf{r}_{i'}^T$.

F. Updating $\mathbf{D}_{C+G+1}$

We update $\mathbf{D}_{C+G+1} = [\mathbf{d}_1^{(C+G+1)}, \ldots, \mathbf{d}_l^{(C+G+1)}, \ldots, \mathbf{d}_{K_{C+G+1}}^{(C+G+1)}] \in R^{d\times K_{C+G+1}}$ atom by atom. By dropping the unrelated terms and defining some notations in (31) of the Appendix, we rewrite the objective function J in (14)

$$\begin{aligned} J\left(\mathbf{d}_l^{(C+G+1)}\right) &= \arg\min_{\mathbf{d}_l^{(C+G+1)}} \sum_{i=1}^{C}\sum_{i\in\mathcal{S}_g} \left\{\left\|\mathbf{W}_{(i,g)} - \mathbf{d}_l^{(C+G+1)}\mathbf{r}_i\right\|_F^2\right\} \\ &2\beta\left\|\mathbf{d}_l^{(C+G+1)}\mathbf{V}_{(C+G+1)}\right\|_F^2 + \sum_{i=1}^{C} \left\{\left\|\mathbf{U}_i - \mathbf{d}_l^{(C+G+1)}\mathbf{r}_i\right\|_F^2\right\}. \end{aligned} \quad (25)$$

Letting $\partial J(\mathbf{d}_l^{(C+G+1)})/\partial\mathbf{d}_l^{(C+G+1)} = 0$, we obtain the updated $\mathbf{d}_l^{(C+G+1)}$ as

$$\mathbf{d}_l^{(C+G+1)} = \frac{1}{2}\left(\mathbf{A} + \beta\mathbf{V}_{C+G+1}\mathbf{V}_{C+G+1}^T\right)^{-1}(\mathbf{B} + \mathbf{R}) \quad (26)$$

where $\mathbf{A} = \sum_{i=1}^{C} \mathbf{r}_i\mathbf{r}_i^T$, $\mathbf{B} = \sum_{i=1, i\neq c}^{C} \mathbf{U}_i\mathbf{r}_i^T$, and $\mathbf{R} = \sum_{i=1}^{C}\sum_{i\in\mathcal{S}_g} \{\mathbf{W}_{(i,g)}\mathbf{r}_i^T\}$.

G. Subdictionary Initialization

Following [35], [36], [38] and [18], we employ K-SVD method [6] to initialize all subdictionaries in Algorithm 1. Specifically, we perform K-SVD on the cth class of training data to initialize $\mathbf{D}_c$ $(c = 1, \ldots, C)$ and perform K-SVD on all the training data to initialize $\mathbf{D}_{C+G+1}$. To simultaneously initialize the FSDs and cluster g $(g = 1, 2, \cdots, G)$, we use the initializing method in [21]. First, we use the K-SVD method over all the training data to get a global dictionary $\mathbf{D}_0 \in \mathbb{R}^{d\times K_0}$ and corresponding coefficients $\mathbf{X}_0 \in \mathbb{R}^{K_0\times N}$. Second, we implement the spectral clustering algorithm on $\mathbf{W} = \mathbf{X}_0^T\mathbf{X}_0 \in \mathbb{R}^{N\times N}$, and group all samples in class c into G clusters. Now, we count the number of samples in each initialized cluster

$$\mathrm{COUNT}(g) = [n(1), \ldots, n(g), \ldots, n(G)] \in \mathbb{R}^G \quad (27)$$

where $n(g)$ is the number of samples in cluster g. And then, the cluster that has the maximum number of samples is the optimizing family cluster $\hat{g}$ of the cth class

$$\hat{g} = \mathrm{argmax}_g \mathrm{COUNT}(g). \quad (28)$$

Finally, the K-SVD method [6] is employed on the gth cluster of training data to initialize $\mathbf{D}_{C+g}$ $(g = 1, \ldots, G)$.

To study the impact of initialization, the random initialization has also been implemented on the face recognition task. We will show the performance comparison with K-SVD initialization and random initialization on four face recognition tasks in our experiments, as shown in Fig. 6. We can see that the improvement of K-SVD initialization in the proposed FDL method is not significant compared with the random initialization.

V. Experiments

We evaluate the performance of the proposed FDL on seven data sets for different basic-level image classification tasks, including four face recognition data sets: Extended YaleB data set [46], Yale face data set [46], Multi-PIE data set [47],

and Aligned Labeled Face in the Wild (aLFW) data set [48]; one handwritten digit recognition data set: USPS data set[3]; one scene classification data set: Scene15 [49]; and one object recognition data set: Caltech101 [50]. To further illustrate the superior performance of the proposed FDL method, some comparison experiments are also conducted on three fine-grained image classification data sets, namely, Oxford Flower 17 data set [51], actions of PPMI [39], and ImageNet subset [52].

A. Classification Schemes

Three classification schemes are chosen in this work as follows.

1) *Global Coding Classifier [42]:* For a testing image $\mathbf{y}$, $\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$. Denote $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_c; \ldots; \mathbf{x}_{C+G+1}]$, and then the final classification based on the reconstruction error $e_c = \|\mathbf{y} - [\mathbf{D}_c, \mathbf{D}_g, \mathbf{D}_{C+G+1}][\mathbf{x}_c; \mathbf{x}_g; \mathbf{x}_{C+G+1}]\|_2^2$ (where $c \in \mathcal{S}_g$) is defined as $c = \arg\min_c e_c$.
2) *Local Coding Classifier [9], [21]:* For a testing image $\mathbf{y}$, the reconstruction error for the cth class is $e_c = \|\mathbf{y} - [\mathbf{D}_c, \mathbf{D}_g, \mathbf{D}_{C+G+1}][\mathbf{x}_c; \mathbf{x}_g; \mathbf{x}_{C+G+1}]\|_2^2 + \lambda\|\mathbf{x}\|_1$ (where $c \in \mathcal{S}_g$). The class identity of $\mathbf{y}$ is $c = \arg\min_c e_c$.
3) *Linear SVMs [35], [38], [49]:* First, we extract the dense SIFT features from images, with the patch size 16 and step size 8. We train an overall dictionary $\mathbf{D}$ on the SIFT features from all categories via the proposed FDL method, and then code the descriptors for each image on $\mathbf{D}$. Second, like in [53], we utilize max pooling technique to pool the coefficients over 4×4, 2×2, and 1×1 subregions, where the coefficients associated with CSDs and FSDs have been adopted, respectively. The intercepted pooled vectors are taken as the representation of the image. For brevity, we called CSD-feature and FSD-feature corresponding to CSDs and FSDs, respectively. Third, we run the proposed FDL ten times with different random splits of training and testing images. Specifically, we first train a coarse classifier on the whole data sets based on FSD-features into super classes, within which a fine classifier is trained based on CSD-features. Linear SVMs are used for both the coarse and the fine classifiers.

For face recognition and handwritten digit recognition tasks, when the number of train samples of each class is relatively small, we use the global coding classifier; otherwise, we adopt the local coding classifier. For scene categorization, object recognition, and fine-grained image classification tasks, linear SVM classification scheme is selected.

B. Face Recognition

To evaluate the proposed FDL method on the face recognition task, we use four real-world benchmarks, as follows.

1) The Extended YaleB data set [18] contains 2414 images of 38 persons, with about 64 images for each person. We resize the original images to the size of 54×48. Then each face image is projected onto a 2592D vector with zero-mean normalization; 20 images per person are randomly selected for training while others for testing.
2) The Yale data set [46] contains 165 images from 15 categories, with 11 face images per person. Each face image is projected onto a 576D vector (pixel resolution) with zero-mean normalization. And then we randomly select six images per category for training and use the rest for testing.
3) The Multi-PIE data set [47] contains 41 368 images of 68 persons. We use five near frontal poses (C05, C07, C09, C27, and C29) for face recognition. Thus, there are 170 images for each individual. Each face image is projected onto a 1024D vector with zero-mean normalization. And then we randomly select six images per category for training and use the rest for testing.
4) The aLFW data set [48] is collected for the unconstrained face recognition. In experiments, 143 persons with no less than 11 images per person are chosen. Ten images per person are randomly selected for training while others for testing. Uniform-LBP features are extracted via dividing a face image into 10×8 patches. Then, the uniform-LBP features are reduced to 1000 dimensions by PCA.

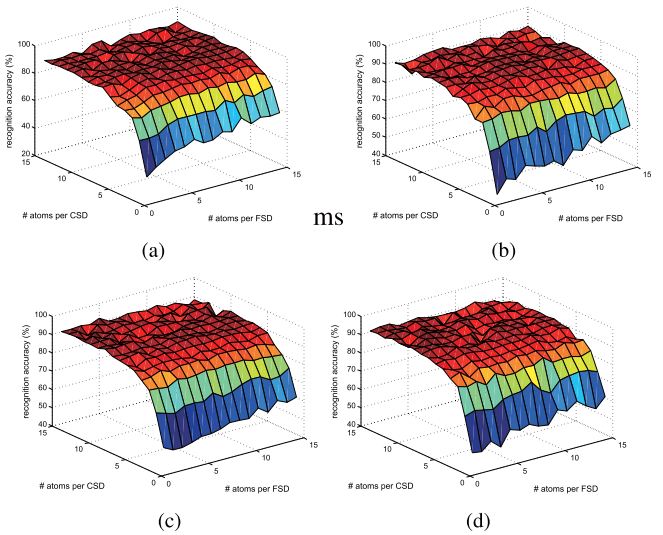


Fig. 3. Accuracy versus dictionary sizes on Extended YaleB data set. (a) #UD atoms = 4. (b) #UD atoms = 6. (c) #UD atoms = 8. (d) #UD atoms = 10.

We compare the proposed FDL method with SRC [42], D-KSVD [31], LC-KSVD [11], DLSI [21], FDDL [18], COPAR [35], and kNN. We use the grid search strategy on CSDs, FSDs, and UD to explore the optimal size of the dictionary for the proposed FDL method.[4] Taking the Extended YaleB data set as an example, the numbers of atoms per CSD, FSD, and UD are set within the range of $\{1, 2, \ldots, 15\}$, $\{1, 2, \ldots, 15\}$, and $\{4, 6, 8, 10\}$. As shown in Fig. 3, we can see that the best performance of the proposed FDL method is achieved when the atom number per CSD, FSD, and UD is 11, 4, and 10, respectively. Compared with the CSD and UD, the classification performance is less sensitive to the number of

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

[4]For a fair comparison, the overall dictionary for the comparative methods is set to the same size in all experiments.

TABLE I
RECOGNITION ACCURACIES (%) ON FOUR FACE DATA SETS

Methods	Datasets			
	Extended-YaleB	Yale	Multi-PIE	aLFW
SRC [42]	90.0	74.6	96.4	71.9
kNN	61.6	43.6	92.1	52.3
D-KSVD [31]	75.3	73.2	91.6	64.5
LS-KSVD [11]	89.5	73.6	95.9	67.3
DLSI [21]	89.0	72.7	96.0	73.0
FDDL [18]	91.7	77.2	97.7	71.4
COPAR [35]	91.5	78.3	98.0	72.9
FDL	**92.8**	**80.2**	**98.9**	**78.6**

(a)

(b)

Fig. 4. Learned dictionary atoms by the proposed FDL method on Yale data set. (a) Class-specific atoms and family-specific atoms. (b) Universal atoms. In (a), four clusters for all face classes are denoted by four color boxes separately. For example, the faces in the red solid line box are class-specific atoms (each column represents one class), while the faces in the red dotted line box are their common family-specific atoms.

the FSD atoms. Because the CSD and UD explicitly code the class-specific patterns and universal patterns, respectively, the FSD implicitly supplements the coding for the some specific patterns, except for the class-specific patterns and universal patterns. The experimental results for these methods on face image data sets are listed in Table I, in which some results are originally taken from [18] and [35]. For a similar atom number of the overall dictionary, the proposed method obtains the best recognition accuracy than other methods.

We also illustrate the learned dictionary atoms of the proposed FDL on the Yale data set in Fig. 4. We can see that the class-specific atoms and family-specific atoms bring in powerful discriminative information. The former brings in the class-discriminative power for each class, while the latter further enhances the cluster-discriminative power for some classes with shared patterns. The universal atoms have no discriminative features, and can only be used to reconstruct universal patterns.

We set appropriate values for the parameters $\{\delta, \beta, \lambda, G\}$ in the proposed FDL method by cross validation over the training set. We set the values of G to 5, 4, 6, and 5 for the

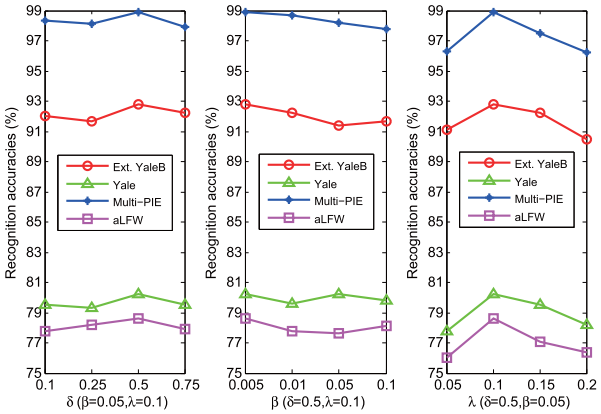


Fig. 5. Sensitivity of parameters δ, β, and λ on Extended-YaleB, Yale, Multi-PIE, and aLFW.

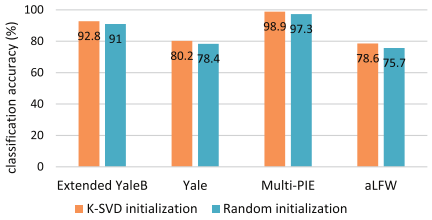


Fig. 6. Comparison with two initializations on four face data sets.

Extended YaleB, Yale, Mylti-PIE, and aLFW data sets, respectively. δ, β, and λ are tuned within the range of {0.1, 0.25, 0.5, 1}, {0.005, 0.01, 0.05, 0.1}, {0.05, 0.1, 0.15, 0.2}, and {0.05, 0.1, 0.15, 0.2} for the Extended YaleB, Yale, Multi-PIE, and aLFW data set, respectively. All the parameters are tuned using fivefold cross validation, and the parameters achieving the best performance are selected. Specifically, we have shown the performance of the proposed method on four face recognition tasks when tuning the tradeoff parameters in Fig. 5. We find that the proposed FDL achieves the best performance when $\{\delta, \beta, \lambda\} = \{0.5, 0.005, 0.1\}$ for all the data sets. Specifically, we can see that the impacts of the parameters δ and β are insensitive to the recognition accuracies of the proposed FDL method. On the other hand, the parameter λ enforces the sparsity of the solution. Although the parameter λ relatively affects the performance, it is within a controllable range. The bigger the λ is, the more sparse the solution will be. Empirically, keeping the sparsity around 10% can yield good results.

C. Handwritten Digit Recognition

We utilize the USPS data set [54] to evaluate the proposed FDL on the handwritten digit recognition task. USPS contains 7291 training images and 2007 testing images. The original image with the size of 16×16 is projected onto a 256D vector. The values of the parameters are set as $\{\delta, \beta, \lambda, G\} = \{0.5, 0.01, 0.1, 1\}$,[5] and the size of the dictionary is $10 \times 28 + 1 \times 9 + 1 \times 8 = 297$ in FDL, where there are ten atoms per CSD, nine atoms per FSD, and eight atoms per UD.

[5]Here, we set the cluster number $G = \{1, \ldots, 4\}$, respectively. The best performance will be achieved when $G = 1$. This is also consistent with the fact that the image in the USPS data set contains only the class-specific patterns and the universal patterns.

TABLE II
RECOGNITION ERROR RATE (%) ON USPS

Methods	Error rate (%)	Methods	Error rate (%)
kNN	5.20	DLSI [21]	3.98
SRSC [55]	6.05	JDL [36]	6.08
SDL-D [8]	3.54	FDDL [18]	3.69
SDL-G [8]	6.67	COPAR [35]	3.61
REC-BL [8]	4.38	TDDL [10]	**2.84**
REC-L [8]	6.83	**FDL**	**3.20**

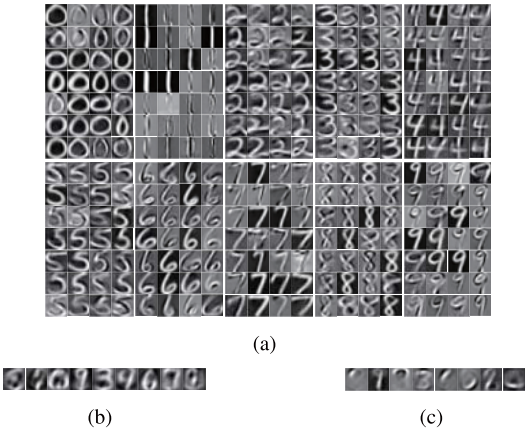


Fig. 7. Learned dictionary atoms of the ten digits by the proposed method on USPS. (b) Family-specific atoms. (c) Universal atoms.

We compare the proposed FDL with SRSC [55], REC-L [8], REC-BL [8], SDL-G [8], SDL-D [8], DLSI [21], TDDL [10], JDL [36], FDDL [18], COPAR [35], and kNN. The sizes of the dictionary for FDDL and COPAR follow those in [18]. The dictionary size of the former is $10 \times 90 = 900$ (90 atoms per class), while the latter is $10 \times 30 + 8 = 308$ (30 atoms per class and eight as the universe). The experimental results are given in Table II, in which some results are originally reported in [18] and [10]. The proposed FDL method achieves the lowest recognition error rate compared with all other methods except for TDDL (2.84%). It is noted that TDDL learns an SVM for each class, and performs classification with a one-versus-all strategy. Compared with TDDL, FDL adopting the global coding classifier of is much simpler.

In addition, compared with dictionary sizes of FDDL and COPAR, FDL uses only 297 atoms in total and can still achieve the leading performance. Fig. 7 shows the learned class-specific atoms, family-specific atoms, and universal atoms for ten digit classes. We can see that the atoms in Fig. 7(a) bring in rich class discriminative information for each class, while the atoms in Fig. 7(b) and (c) have no class-discriminative information and can be used only to construct the nonclass-specific patterns in the data set.

D. Object Recognition

We evaluate the proposed FDL on the Caltech101 data set, which contains 9144 images from 101 classes and one background class. The image number per category varies from 31 to 800. Following the common experimental settings, we partition the whole data set into training images (5, 10, 15, 20, 25, and 30 images per class) and testing images (no more than 50 images per class). Usually, the size of the dictionary is set as $K = 2048$. Therefore, we set the dictionary size to $102 \times 16 + 8 \times 27 + 200 = 2048$, where the atom number per CSD, FSD, and UD is 16, 27, and 200, respectively, and the cluster number of classes is 8. The values of the parameters are set as $\{\delta, \beta, \lambda, G\} = \{0.1, 0.05, 0.15, 8\}$.

TABLE III
CLASSIFICATION ACCURACIES (%) ON CALTECH101

Methods	Training samples per class					
	5	10	15	20	25	30
SPM [49]	-	-	56.4	-	-	64.6
ScSPM [53]	-	-	67.0	-	-	73.2
Griffin *et al.* [62]	44.2	54.5	59.0	63.3	65.8	67.6
LLC [23]	51.15	59.77	65.43	67.74	70.16	73.44
Boureau *et al.* [63]	-	-	-	-	-	77.3
LSC [59]	-	-	-	-	-	74.2
Duchenne *et al.* [64]	-	-	75.3	-	-	80.1
Feng *et al.* [66]	-	-	70.3	-	-	82.6
Goh *et al.* [67]	-	-	71.1	-	-	78.9
MKL-LDE [60]	59.2	68.9	74.9	77.2	79.2	-
LC-KSVD [11]	54.0	63.1	67.7	70.5	72.3	73.6
CRBM [61]	56.7	66.7	71.3	74.2	76.2	77.8
KC [56]	-	-	-	-	-	64.2
SRC [42]	48.8	60.1	64.9	67.7	69.2	70.7
K-SVD [6]	49.8	59.8	65.2	68.7	71.0	73.2
D-KSVD [31]	49.6	59.5	65.1	68.6	71.1	73.0
Nguyen *et al.* [58]	56.5	67.2	72.5	75.8	77.6	80.1
MKSR [57]	59.9	69.5	75.7	79.7	80.8	82.9
DeCAF-fc6 [65]	-	-	-	-	-	**86.9**
COPAR [35]	60.4	71.3	75.1	77.6	80.5	83.3
FDL without FSD	61.6	73.2	80.9	81.4	82.0	82.9
FDL	**63.7**	**75.5**	**81.4**	**83.2**	**83.6**	84.2

We compare the proposed FDL with KC [56], SRC [42], K-SVD [6], D-KSVD [31], LC-KSVD [11], COPAR [35], MKSR [57], Nguyen *et al.* [58], ScSPM [53], spatial pyramid matching (SPM) [49], LLC [23], Laplacian sparse coding (LSC) [59], MKL-LDE [60], CRBM [61], Griffin *et al.* [62], Boureau *et al.* [63], and Duchenne *et al.* [64]. Moreover, as the state-of-the-art methods, convolutional neural networks (CNNs) [65] are also used for comparison. We repeat the experiments ten times with the random splits of the training and testing images. The average per-class recognition accuracy is recorded for each run. We report the average recognition accuracy in Table III, in which some results are originally reported in [11] and [57]. We can find that FDL and COPAR achieve higher recognition rates than other dictionary learning methods, since they pursue pure class-specific representations. Compared with COPAR, the proposed method incorporates supervised learning for the overall dictionary and unsupervised learning for CSDs, FSDs, and UD, and thus outperforms COPAR. In addition, the recognition rate of the proposed method exceeds 80% when the number of training images per class is 15. Note that CNN (i.e., DeCAF-fc6 in Table III) achieves the amazing performance by extracting the features on FC6 layers on a CNN, which is pretrained well on the ImageNet database [52]. The proposed method is trained on a smaller data set, which cannot train CNN well. Therefore, the proposed method can use a relatively small number of training data to learn a pure CSD.

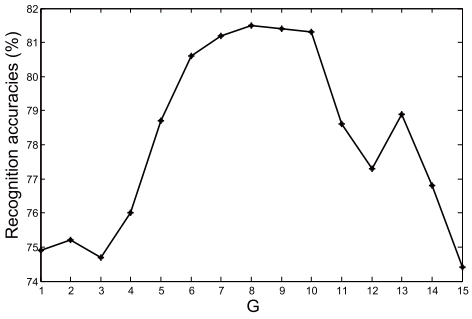


Fig. 8. Recognition accuracies (%) of FDL by varying the number of clusters on Caltech101.

Fig. 9. Some example images of 12 categories from ImageNet database.

TABLE IV
ACCURACIES (%) ON SCENE15

Methods	Accuracies (%)	Methods	Accuracies (%)
ScSPM [53]	80.28	K-SVD [6]	80.0
KSPM [49]	81.40	LC-KSVD [11]	85.4
LScSPM [68]	**89.75**	MMDL [29]	78.1
LLC [23]	79.24	KC [56]	76.67
Boureau *et al.* [63]	83.50	COPAR [35]	86.37
FDL	**88.48**	**FDL+LSC**	**92.64**

To illustrate the importance of family-specific atoms, we also compare the proposed FDL method without universal atoms on Caltech 101 data set, and we can see that our performance is decreased by about 1∼2%, but is higher than COPAR, which is equipped with the universal atoms. This illustrates that the combination of family-specific atoms and universal atoms is more effective than either family-specific atoms or universal atoms. In particular, family-specific atoms are more suitable than universal atoms, which can be seen the family-specific atoms with only one cluster.

To study the sensitivity of the cluster number G, we also conduct additional experiments on the Caltech 101 data set by varying the values of G from 1 to 15. Fig. 8 presents the curve of accuracy versus G. We can see that the proposed FDL achieves the best performance on Caltech 101 when $G = 8$. We can see that the performance of the proposed FDL is insensitive to the number of clusters (G) within in the range $[7, 10]$.

E. Scene Categorization

As one of the most complete scene categorization data sets, the Scene15 data set contains 15 scene categories and 4485 images in total, with each category containing from 200 to 400 images. The average image size is 300×250. The 15 categories include living rooms, streets, kitchens, and so on. We randomly select 100 images per class for training and use the rest for testing. The values of the parameters are set as $\{\delta, \beta, \lambda, G\} = \{0.1, 0.05, 0.15, 4\}$. For FDL, the size of the overall dictionary is $15 \times 50 + 4 \times 30 + 154 = 1024$, in which 50 is the size of each CSD, 30 is the size of each FSD, and 154 is the size of UD.

The comparative methods include some popular methods: ScSPM [53], KSPM [49], LScSPM [68], LLC [23], and the method of Boureau *et al.* [63], and the representative methods related to dictionary learning: KC [56], K-SVD [6], LC-KSVD [11], MMDL [29], as well as COPAR [35]. The comparison results are shown in Table IV, in which some results are reported in [68] and [63]. We can see that the proposed FDL method outperforms all other methods except LScSPM [68]. It is worth noting that LScSPM adopts an extra assumption of the manifold structure over the data samples, and takes LSC to preserve the consistence in sparse representation of similar local features. For a fair comparison with LScSPM, we also extend FDL to incorporate LSC, called FDL+LSC, and find that FDL+LSC is better than LScSPM by about 4%.

F. Fine-Grained Image Categorization

The above experiments are to evaluate the proposed FDL for basic-level image classification tasks. To further validate the performance of FDL, we compare it with other dictionary learning methods for some fine-grained image classification tasks. Three fine-grained image data sets are used in experiments.

1) The ImageNet subset [52] contains 12 outdoor game categories including "cricket" (n00476389), "baseball" (n00471613), "American football" (n00469651), "polo" (n00477639), "hurling" (n00470830), "soccer" (n00478262), "tennis" (n00483205), "horse racing" (n00450070), "netball" (n00482122), "golf" (n00466273), "croquet" (n00466880), and "frisbee" (n03397947). This data set contains 6000 images, with about 500 images per category. All images are resized with the maximum width or height of 300 pixels while keeping the fixed aspect ratio; 50 images per category are randomly selected for training while others for testing. Fig. 9 shows some examples in the ImageNet subset. We can see that all image classes are visually similar.
2) The Oxford Flower 17 data set [51] contains 1360 images from 17 categories, where each category includes 80 images. We use all the images without bounding box provided by the data set. We use the standard train/test split provided by the data set, in which 40 images are used as training samples. Like [53], all images are resized with the maximum side of 300 pixels while keeping the fixed aspect ratio.

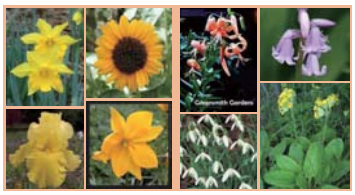


Fig. 10. Some samples in the Oxford Flower 17 data set. Each image represents one category.

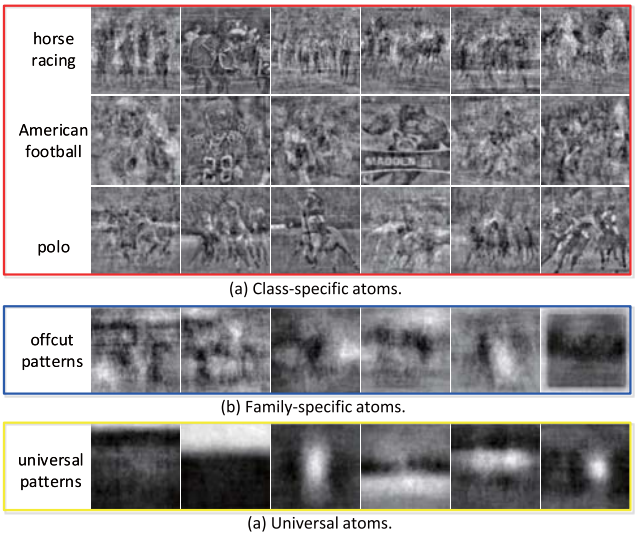


Fig. 11. Visualization of dictionary atoms by the proposed method on the ImageNet subset.

Fig. 10 shows some examples in Oxford Flower 17, where some categories have higher visual similarities.

3) The PPMI data set [39] contains images of human interacting with 12 different musical instruments, including "bassoon," "cello," "clarinet," "erhu," "flute," "French horn," "guitar," "harp," "recorder," "saxophone," "trumpet," and "violin". Based on the interactions between humans and instruments, PPMI can be grouped into 24 classes. Following [39], we use the normalized images with size 256×256 for feature extraction, and take 100 images per class as training images and the remaining 100 images for testing.

Similar to the basic-level classification tasks, we set $\{\delta, \beta, \lambda, G\} = \{0.1, 0.05, 0.15, 4\}$, $\{\delta, \beta, \lambda, G\} = \{0.5, 0.01, 0.15, 3\}$, and $\{\delta, \beta, \lambda, G\} = \{0.1, 0.05, 0.1, 12\}$ for ImageNet subset, Oxford flower 17, and PPMI, respectively. The dictionary size of FDL is set as $12 \times 30 + 4 \times 40 + 60 = 580$ for ImageNet subset, $17 \times 25 + 3 \times 25 + 100 = 600$ for Oxford Flower 17, and $24 \times 45 + 12 \times 50 + 150 = 1830$ for PPMI.

1) Results on the ImageNet Subset: We compare the proposed FDL with methods based on SPM (i.e., ScSPM [53] and LLC [23]) and some dictionary learning methods related to our work (i.e., D-KSVD [31], FDDL [18], JDL [36], CSDL [38], and COPAR [35]). The number of dictionary atoms in FDDL is set as the number of training samples, while the number of commonly shared dictionary atoms and the number of CSD atoms in JDL, CSDL, and COPAR are set to 45 and 50, respectively. Fig. 11 visualizes some learned atoms by the proposed FDL. We can see that the class-specific atoms reflect the class-discriminative patterns. Family-specific atoms capture some shared patterns in the three classes, i.e., "horse racing," "American football," and "polo." Universal atoms cover the outlines of background in this data set. Table V lists the classification accuracies of various methods. FDL outperforms the other comparative methods, with significant improvements over the second best one. Although each class of ImageNet is highly relevant in visual similarities, FDL focusing on the fine-grained dictionaries can represent their similar yet different-specific patterns by the corresponding different-specific dictionaries. We show the confusion matrix on the data set in Fig. 12.

TABLE V
CATEGORIZATION ACCURACIES (%) ON THE IMAGENET SUBSET

Methods	Accuracies (%)	Methods	Accuracies (%)
ScSPM [53]	38.87 ± 1.0	JDL [36]	40.31 ± 1.3
D-KSVD [31]	34.43 ± 3.1	CSDL [38]	43.60 ± 1.9
LLC [23]	35.92 ± 0.9	COPAR [35]	42.91 ± 2.2
FDDL [18]	38.52 ± 2.7	**FDL**	**51.24 ± 2.1**

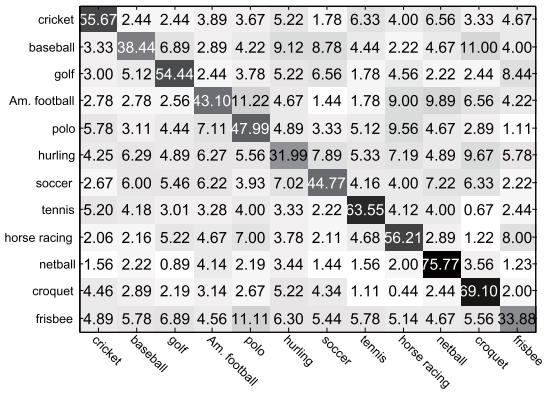


Fig. 12. Confusion matrix (%) of the proposed FDL on the ImageNet subset.

2) Results on the Oxford Flower 17 Data Set: We compare the proposed FDL with some baselines (i.e., ScSPM [53], LLC [23], and SRC) and some dictionary learning methods, i.e., FDDL [18], JDL [36], CSDL [38], and COPAR [35]. The dictionary sizes of JDL, CSDL, and COPAR are set to $17 \times 30 + 80 = 590$, where 30 and 80 denote the number of CSD atoms and commonly shared dictionary atoms, respectively. The size of the dictionary in FDDL is set to the size of the training set. The comparison results are listed in Table VI, in which some results are reported in [69]. Again, FDL achieves the remarkable improvements over most of the comparative methods. For the similar patterns among some categories, the proposed FDL can further exclude them from the class-specific patterns, which makes these class-specific patterns be high discriminative.

3) Results on the PPMI Data Set: We compare the proposed FDL with some baselines (i.e., SPM [53] and LLC [23], and Grouplet [39]) and some dictionary learning methods (i.e., FDDL [18], JDL [36], CSDL [38], and COPAR [35]). Similarly, the dictionary sizes of JDL, CSDL, and COPAR are set to $24 \times 80 + 150 = 2070$, where 80 and 150 denote

TABLE VI
CATEGORIZATION ACCURACIES (%) ON OXFORD FLOWER 17

Methods	Accuracies (%)	Methods	Accuracies (%)
ScSPM [53]	64.2 ± 1.3	JDL [36]	48.4 ± 1.1
SRC [42]	62.3 ± 2.6	CSDL [38]	71.43 ± 1.7
LLC [23]	45.00 ± 2.1	COPAR [35]	62.6 ± 1.8
FDDL [18]	63.68 ± 2.4	**FDL**	**77.16 ± 2.2**

TABLE VII
CATEGORIZATION ACCURACIES (%) ON PPMI

Methods	Accuracies (%)	Methods	Accuracies (%)
ScSPM [53]	35.47 ± 0.9	SPM [49]	33.24 ± 0.6
LLC [42]	30.58 ± 1.7	CSDL [38]	43.11 ± 2.7
FDDL [18]	42.15 ± 1.8	COPAR [35]	42.68 ± 2.5
Grouplet [39]	37.10 ± 0.7	**FDL**	**49.36 ± 2.0**

the number of CSD atoms and commonly shared dictionary atoms, while the dictionary size of FDDL is equipped with the same size of the training set. Table VII lists the experimental results of various methods. We can see that FDL outperforms the comparative methods. Compared with the second best one (CSDL with 43.11%), the proposed FDL has gained 5.68% improvement. The proposed FDL is more suitable for addressing the two highly similar classes in this data set, i.e., class "play instrument" and "with instrument." Moreover, to show the important role of the atoms from UD, CSD, and FSD, we also visualize the atoms. The visualized results of a part of atoms are shown in Fig. 1. We can see that the atoms from CSDs are related to the image labels, such as action of play guitar versus class label "play guitar." We can see that the atoms corresponding to class "play guitar" bring in more obvious interactive actions than the atoms corresponding to class "with guitar." In the shared atoms corresponding to class "play guitar" and class "with guitar," there are abstractly shared patterns between them. The atoms of UD have the common background, such as human body and scene.

VI. TIME COMPLEXITY

We analyze the time complexity in the training phase. First, updating the coding coefficient for each sample is a traditional sparse coding problem. The corresponding time complexity for each image is approximately $O(d^2K^\varepsilon)$ by following [69], where $\varepsilon \geq 1.2$ is a constant, d is the feature dimensionality of the sample, and K is the number of dictionary atoms. Then, the time complexity of all N samples is $O(Nd^2K^\varepsilon)$. Second, the time complexity of updating clusters in (17) is $\sum_{c=1}^{C}\sum_{g=1}^{G} O(dK_cN_{C+g})$. Third, the time complexity of updating dictionary atoms is $\sum_{c=1}^{C}(O(dK_cN_c)+O(dK_cN_{C+g})+O(dK_cN_C))$. Therefore, the overall time complexity of FDL is approximately $\vartheta(O(Nd^2K^\varepsilon)+\sum_{c=1}^{C}(O(dK_cN_c)+O(dK_cN_{C+g})+O(dK_cN_C)+\sum_{g=1}^{G}O(dK_cN_{C+g})))$, where ϑ is the number of iterations. In this paper, the experiments are implemented in MATLAB environment running in DELL Server with 32-GB memory. Taking Caltech101 data set as an example, the training time of the proposed FDL method is about 4.3 h. Specifically, we compare FDL with SRC and COPAR in terms of the computation time for classifying one testing image. With the same size of the dictionary, the testing times for one image of SRC, COPAR, and FDL are 704.6, 815.8, and 897.5 s, respectively. However, the accuracy of FDL (84.2%) is higher than SRC (70.7%) and COPAR (83.3%). By jointly considering the accuracy and the testing time, the proposed method is more competitive than SRC and COPAR.

VII. CONCLUSION

In this paper, to obtain the high-quality representation of images for various kinds of classification tasks, we propose a novel FDL method by simultaneously learning three types of specific dictionaries, including CSDs, FSDs, and a UD. We evaluate the proposed FDL method by conducting extensive experiments on ten data sets covering various classification tasks. The proposed FDL method achieves the promising performance in comparison with other related methods.

APPENDIX
SOME DEFINITIONS

Some definitions in this paper are as follows:

$$\begin{cases} \mathbf{r}_i = \mathbf{e}_l^c \mathbf{H}_c \mathbf{X}_i \in R^{1\times N_i}, \quad \text{for } \forall i, i = 1,\ldots,C \\ \mathbf{U}_i = \mathbf{Y}_i - \mathbf{D}\mathbf{H}_{/c}^T \mathbf{H}_{/c}\mathbf{X}_i - \mathbf{d}\tilde{\mathbf{H}}_c\mathbf{X}_i, \quad \text{for } \forall i, i = 1,\ldots,C \\ \mathbf{V}_c = \mathbf{D}\mathbf{H}_{/(c)}^T \\ \mathbf{Z}_c = \mathbf{Y}_c - \mathbf{D}\big(\mathbf{H}_{C+g}^T \widehat{\mathbf{H}_{C+g}} + \mathbf{H}_{C+G+1}^T \mathbf{H}_{C+G+1}\big)\mathbf{X}_c \\ \qquad - \tilde{\mathbf{d}}_c \mathbf{H}_c \mathbf{X}_c. \end{cases} \tag{29}$$

$$\begin{cases} \mathbf{r}_i = \mathbf{e}_l^{C+g} \mathbf{H}_{C+g}\mathbf{X}_i \in R^{1\times N_i}, \quad \text{for } \forall i, i = 1,\ldots,C \\ \mathbf{U}_i = \mathbf{Y}_i - \mathbf{D}\mathbf{H}_{/c}^T \mathbf{H}_{/c}\mathbf{X}_i - \mathbf{d}\tilde{\mathbf{H}}_c\mathbf{X}_i, \quad \text{for } \forall i, i = 1,\ldots,C \\ \mathbf{V}_{C+g} = \mathbf{D}\mathbf{H}_{/C+g}^T \\ \mathbf{Z}_{i'} = \mathbf{Y}_{i'} - \mathbf{D}\big(\mathbf{H}_{i'}^T \mathbf{H}_{i'} + \mathbf{H}_{C+G+1}^T \mathbf{H}_{C+G+1}\big)\mathbf{X}_{i'} \\ \qquad - \mathbf{d}\widehat{\tilde{\mathbf{H}}_{C+g}}\mathbf{X}_{i'}, \quad \text{for } \forall i', i' = 1,\ldots,C. \end{cases} \tag{30}$$

$$\begin{cases} \mathbf{r}_i = \mathbf{e}_l^{C+G+1} \mathbf{H}_{C+G+1}\mathbf{X}_i \in R^{1\times N_i}, \ \text{for } \forall i, i = 1,\ldots,C, \\ \mathbf{U}_i = \mathbf{Y}_i - \mathbf{D}\mathbf{H}_{/C+G+1}^T \mathbf{H}_{/C+G+1}\mathbf{X}_i \\ \qquad - \mathbf{d}\tilde{\mathbf{H}}_{C+G+1}\mathbf{X}_i, \quad \text{for } \forall i, i = 1,\ldots,C, \\ \mathbf{V}_{C+G+1} = \mathbf{D}\mathbf{H}_{/(C+G+1)}^T, \\ \tilde{\mathbf{d}}_{C+G+1} = \left(\sum_{m=1,m\neq l}^{K_{C+G+1}} \mathbf{d}_m^{(C+G+1)} \mathbf{e}_m^{C+G+1}\right) \\ \mathbf{W}_{(i,g)} = \mathbf{Y}_i - \mathbf{D}\big(\mathbf{H}_i^T \mathbf{H}_i + \mathbf{H}_{C+g}^T \mathbf{H}_{C+g}\big)\mathbf{X}_i \\ \qquad - \mathbf{d}\tilde{\mathbf{H}}_{C+G+1}\mathbf{X}_i, \quad \text{for } \forall i, i = 1,\ldots,C. \end{cases} \tag{31}$$

REFERENCES

[1] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying discriminative visual codebook generation with classifier training for object category recognition," in *Proc. CVPR*, Jun. 2008, pp. 1–8.
[2] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1243–1256, Jul. 2008.
[3] Y. Sun, X. Tao, Y. Li, and J. Lu, "Dictionary learning for image coding based on multisample sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 2004–2010, Nov. 2014.

[4] H. Xiong, Z. Pan, X. Ye, and C. W. Chen, "Sparse spatio-temporal representation with adaptive regularized dictionary learning for low bit-rate video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 710–728, Apr. 2013.

[5] M. Xu, S. Li, J. Lu, and W. Zhu, "Compressibility constrained sparse representation with learnt dictionary for low bit-rate image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1743–1757, Oct. 2014.

[6] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[7] M. Zhou, H. Chen, L. Ren, G. Sapiro, L. Carin, and J. W. Paisley, "Non-parametric Bayesian dictionary learning for sparse image representations," in *Proc. NIPS*, 2009, pp. 2295–2303.

[8] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Proc. NIPS*, 2008, pp. 1033–1040.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. CVPR*, 2008, pp. 1–8.

[10] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[11] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. CVPR*, 2011, pp. 1697–1704.

[12] Z. Jiang, G. Zhang, and L. S. Davis, "Submodular dictionary learning for sparse coding," in *Proc. CVPR*, 2012, pp. 3418–3425.

[13] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

[14] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1576–1588, Aug. 2012.

[15] A. Castrodad and G. Sapiro, "Sparse modeling of human actions from motion imagery," *Int. J. Comput. Vis.*, vol. 100, no. 1, pp. 1–15, 2012.

[16] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," Tech. Rep., 2008.

[17] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. ECCV*, 2010, pp. 448–461.

[18] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. ICCV*, 2011, pp. 543–550.

[19] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," in *Proc. ICCV*, Nov. 2011, pp. 707–714.

[20] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognit.*, vol. 45, no. 11, pp. 3902–3911, 2012.

[21] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. CVPR*, 2010, pp. 3501–3508.

[22] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain, "Image annotation by *k*NN-sparse graph-based label propagation over noisily tagged Web images," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, 2011, Art. no. 14.

[23] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. CVPR*, 2010, pp. 3360–3367.

[24] R. Jenatton, J. Mairal, F. Bach, and G. Obozinski, "Proximal methods for sparse hierarchical dictionary learning," in *Proc. ICML*, 2010, pp. 487–494.

[25] W. Zhang, A. Surve, X. Fern, and T. Dietterich, "Learning non-redundant codebooks for classifying complex objects," in *Proc. ICML*, 2009, pp. 1241–1248.

[26] Y.-T. Chi, M. Ali, A. Rajwade, and J. Ho, "Block and group regularized sparse modeling for dictionary learning," in *Proc. CVPR*, 2013, pp. 377–382.

[27] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, Sep. 2011.

[28] M. Jian and C. Jung, "Class-discriminative kernel sparse representation-based classification using multi-objective optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 18, pp. 4416–4427, Sep. 2013.

[29] X.-C. Lian, Z. Li, B.-L. Lu, and L. Zhang, "Max-margin dictionary learning for multiclass image categorization," in *Proc. ECCV*, 2010, pp. 157–170.

[30] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. CVPR*, 2010, pp. 3517–3524.

[31] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. CVPR*, 2010, pp. 2691–2698.

[32] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. CVPR*, 2013, pp. 676–683.

[33] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[34] L. Shen, S. Wang, G. Sun, S. Jiang, and Q. Huang, "Multi-level discriminative dictionary learning towards hierarchical visual categorization," in *Proc. CVPR*, 2013, pp. 383–390.

[35] S. Kong and D. Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," in *Proc. ECCV*, 2012, pp. 186–199.

[36] N. Zhou, Y. Shen, J. Peng, and J. Fan, "Learning inter-related visual dictionary for object recognition," in *Proc. CVPR*, 2012, pp. 3490–3497.

[37] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *Proc. CVPR*, 2011, pp. 1577–1584.

[38] S. Gao, I. W.-H. Tsang, and Y. Ma, "Learning category-specific dictionary and shared dictionary for fine-grained image categorization," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 623–634, Feb. 2013.

[39] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. CVPR*, 2010, pp. 9–16.

[40] K. Li, J. Yang, and J. Jiang, "Nonrigid structure from motion via sparse representation," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1401–1413, Aug. 2015.

[41] J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua, "Semantic-gap-oriented active learning for multilabel image annotation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2354–2360, Apr. 2012.

[42] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[43] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1395–1408, Jul. 2009.

[44] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5343–5355, Dec. 2015.

[45] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[46] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[47] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[48] L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," in *Proc. ACCV*, 2010, pp. 88–97.

[49] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 2169–2178.

[50] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 59–70, Jan. 2007.

[51] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[53] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. CVPR*, 2009, pp. 1794–1801.

[54] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "USPS handwritten digit database," *Image Vis. Comput.*, to be published.

[55] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. NIPS*, 2006, pp. 609–616.
[56] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Proc. ECCV*, 2008, pp. 696–709.
[57] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Multiple kernel sparse representations for supervised and unsupervised learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2905–2915, Jul. 2014.
[58] H. Van Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Design of non-linear kernel dictionaries for object recognition," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5123–5135, Dec. 2013.
[59] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Proc. ICCV*, 2011, pp. 2486–2493.
[60] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.
[61] K. Sohn, D. Y. Jung, H. Lee, and A. O. Hero, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition," in *Proc. ICCV*, 2011, pp. 2643–2650.
[62] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep., 2007.
[63] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. CVPR*, 2010, pp. 2559–2566.
[64] O. Duchenne, A. Joulin, and J. Ponce, "A graph-matching kernel for object categorization," in *Proc. ICCV*, 2011, pp. 1792–1799.
[65] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014, pp. 647–655.
[66] J. Feng, B. Ni, Q. Tian, and S. Yan, "Geometric ℓ_p-norm feature pooling for image classification," in *Proc. CVPR*, 2011, pp. 2609–2704.
[67] H. Goh, N. Thome, M. Cord, and J.-H. Lim, "Unsupervised and supervised visual codes with restricted Boltzmann machines," in *Proc. ECCV*, 2012, pp. 298–311.
[68] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely—Laplacian sparse coding for image classification," in *Proc. CVPR*, 2010, pp. 3555–3561.
[69] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, 2014.

Xiangbo Shu received the Ph.D. degree from Nanjing University of Science and Technology, Nanjing, China, in 2016.

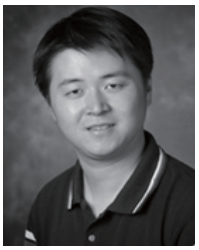
From 2014 to 2015, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is currently an Assistant Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include computer vision and machine learning.

Mr. Shu was a recipient of the Best Student Paper Awards in MMM 2016 and the Best Paper Finalist in ACM MM 2015.

Jinhui Tang received the B.E. and Ph.D. degrees from University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively.

From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. He has authored over 100 journal and conference papers. His research interests include large-scale multimedia search.

Prof. Tang was a recipient of the ACM China Rising Star Award and a co-recipient of the Best Student Paper Award in MMM 2016, and the Best Paper Award in ACM MM 2007, PCM 2011, and ICIMCS 2011.



Guo-Jun Qi received the Ph.D. degree from the Department of Electrical and Computer Engineering, Beckman Institute, University of Illinois at Urbana–Champaign, Urbana, IL, USA, in 2013.

He is currently an Assistant Processor with the Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL, USA. He was with the Department of Electrical and Computer Engineering, Beckman Institute, University of Illinois at Urbana–Champaign. His current research interests include pattern recognition, machine learning, computer vision, and multimedia.

Mr. Qi was a recipient of the Best Paper Award in ACM MM 2007.

Zechao Li received the the B.E. degree from University of Science and Technology of China, Hefei, China, in 2008 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2013.

He is an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include large-scale multimedia understanding, social media mining, and subspace learning.

Mr. Li was a recipient of the 2015 Excellent Doctoral Dissertation of the Chinese Academy of Sciences, the 2015 Excellent Doctoral Dissertation of the China Computer Federation, the Top 10% Paper Award of the IEEE MMSP 2015, and the 2013 President Scholarship of the Chinese Academy of Sciences.

Yu-Gang Jiang received the Ph.D. degree in computer science from City University of Hong Kong, Hong Kong, in 2009.

From 2008 to 2011, he was with the Department of Electrical Engineering, Columbia University, New York City, NY, USA. He is currently a Professor of Computer Science with Fudan University, Shanghai, China. His research interests include multimedia retrieval and computer vision.

Shuicheng Yan is an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. He has authored or co-authored hundreds of technical papers, with a Google Scholar citation of more than 15 000 times. He has an H-index of 52. His research interests include machine learning, computer vision, and multimedia.

Dr. Yan was a recipient of the Best Paper Awards from ACM MM 2013 (Best Paper and Best Student Paper awards), MMM 2016 (Best Student Paper), ACM MM 2012 (Best Demo), PCM 2011, ACM MM 2010, ICME 2010, and ICIMCS 2009; the winner prizes of the classification task in the PASCAL VOC 2010-2012; the 2011 Singapore Young Scientist Award; and the 2012 NUS Young Researcher Award. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *ACM Transactions on Intelligent Systems and Technology*.