Contextual Online Dictionary Learning for Hyperspectral Image Classification

Wei Fu, *Student Member, IEEE*, Shutao Li, *Senior Member, IEEE*, Leyuan Fang, *Senior Member, IEEE*, and Jón Atli Benediktsson, *Fellow, IEEE*

***Abstract*— Sparse representation (SR) has been successfully used in the classification of hyperspectral images (HSIs) by representing HSI pixels over a dictionary and yielding discriminative sparse coefficients. Most of SR-based classification methods construct the dictionary by directly using some labeled pixels as atoms. Such dictionary can lead to inefficient SR for large-sized HSIs, and may be incomplete when the number of labeled pixels is less than the number of spectral bands. This paper proposes a contextual online dictionary learning (DL) method for HSIs classification, which learns a dictionary over the whole image rather than few labeled pixels. The proposed method can effectively and efficiently improve the adaptive representation capability of different pixels with an online learning mechanism. Specifically, the contextual characteristics of the HSI are integrated with discriminative spectral information for online DL, i.e., pushing similar pixels in neighborhood to share similar sparse coefficients with respect to the well-learned dictionary. By this way, the obtained sparse coefficients are structured and discriminative. Finally, a traditional classifier, i.e., the linear support vector machine, is applied to the sparse coefficients, and the final classification results are obtained. Experimental results on real HSIs show the effectiveness of the proposed method.**

***Index Terms*— Classification, contextual characteristics, hyperspectral images (HSIs), online dictionary learning (DL), sparse representation (SR).**

I. Introduction

CLASSIFICATION of hyperspectral images (HSIs) can be used in agriculture [1], [2], military [3], [4], and environmental monitoring [5], [6]. Therefore, HSI classification is an important research field in remote sensing. Especially, supervised classification, which categorizes test pixels with a classifier trained from some labeled pixels (i.e., training pixels), gradually becomes a hot topic in recent years.

Manuscript received March 21, 2017; revised July 17, 2017 and September 21, 2017; accepted October 6, 2017. Date of publication October 27, 2017; date of current version February 27, 2018. This work was supported by the National Natural Science Fund of China for Distinguished Young Scholars under Grant 61325007, by the National Natural Science Fund of China for International Cooperation and Exchanges under Grant 61520106001, and by the Fund of Hunan Province for Science and Technology Plan Project under Grant 2017RS3024. *(Corresponding author: Shutao Li.)*

W. Fu, S. Li, and L. Fang are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: fuweihunandaxue@gmail.com; shutao_li@hnu.edu.cn; fangleyuan@gmail.com).

J. A. Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavk, Iceland (e-mail: benedikt@hi.is).

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/TGRS.2017.2761893

Previously, some classical pixelwise classification methods [7]–[10] are developed by only using the discriminative information of spectral signatures. Without considering the contextual information of pixels, those pixelwise methods usually result in not very high classification accuracy. Many spectral–spatial classification methods [11]–[18] are then proposed to exploit both the contextual information and the spectral information, and achieve outstanding performance. For instance, Li *et al.* [13] extract the spectral–spatial features with the local binary patterns (LBP) [14] and then input those features into the extreme learning machine (ELM) [15] classifier for classification. Besides, Fang *et al.* [19] extract the spectral–spatial features using superpixel segmentation and multiple kernels construction, and then achieve classification results by the support vector machine (SVM) classifier [7].

Recently, sparse representation (SR)-based [20] spectral–spatial classification methods [21]–[26] become popular. In the SR model, spectral signatures of pixels are represented with the linear combination of only a few atoms in a dictionary. Most SR-based HSI classification methods [23]–[25] regard training samples (i.e., few selected pixels with known class labels) as atoms of the dictionary. It is convenient to determine the class labels of test pixels according to the labels of selected atoms for representation. However, such dictionary seriously depends on the selection of training pixels and sometimes will be impractical. That is, too large number of training pixels means a very large dictionary, which can lead to inefficient computation. On the other hand, very few training pixels indicate a very small dictionary, which may not sparsely represent other pixels and thus degrade the performance of the SR model.

To overcome the drawbacks of simply using training pixels as atoms, some dictionary learning (DL) methods [27]–[29] are proposed to learn a compact and discriminative dictionary from training pixels. Due to the use of pixels' label information in DL, those methods are recognized as supervised DL methods. Although supervised DL methods can improve performances in terms of computational efficiency and classification accuracy, they are also influenced by the selection of training pixels. Once training pixels are badly chosen, it is not ensured that the learned dictionary can sparsely and discriminatively represent test pixels, or even suffering from a risk of overtraining. As a consequence, the performance of SR-based classification can be degraded.

Unlike supervised DL methods, which learn a dictionary using training pixels, unsupervised DL methods, e.g., spatial-aware DL (SADL) method [30], learn the dictionary from the whole data set without considering label information. Moreover, it is convenient to incorporate prior information into dictionary by using suitable sparsity inducing regularity [30] and result in discriminative sparse coefficients for classification [23], [25]. Considering these advantages, we focus on using the whole data set to learn the dictionary, which is the same as unsupervised DL methods.

The traditional DL mechanism [30]–[32] attempts to accurately minimize an empirical cost function to calculate dictionary and sparse coefficients. To serve this purpose, the sparse coding of all data and dictionary update is usually alternated for dozens of iterations. In practice, a remote sensing HSI usually contains a large number of pixels, which significantly increase the number of input data in DL. The iterative sparse coding of all data during DL can lead to high computation cost. Therefore, how to efficiently learn the dictionary is an important issue that we need to solve.

Fortunately, as pointed out by Bottou and Bousquet [33], the minimization of an expected cost function is more interesting than the minimization of the empirical cost function. To efficiently minimize the expected cost function, the online learning mechanism [34], which inputs a few new pixels into SR learning in each iteration, is considered. In addition, HSI pixels' contextual information, which can efficiently improve classification accuracy [35]–[37], is incorporated into online DL. That is, the spectral similarity of neighboring pixels is calculated and used as a prior knowledge to induce structural sparse coefficients. The obtained sparse coefficients can be classified with linear classifier, e.g., the linear SVM. To this end, a new DL method called contextual online DL (CODL) is carried out in this paper. Experimental results show that the proposed CODL method outperforms some classic or state-of-the-art HSIs classification methods.

The remainder of this paper is structured as follows. Some related works are introduced in Section II. The proposed CODL method is described in Section III. Section IV presents the experimental results and discussions. Finally, Section V summarizes this paper and gives some suggestions about future works.

II. Related Work

In this section, we briefly introduce unsupervised DL for HSIs. As shown in [38], [39], HSI pixels can be approximately represented by the multiply of sparse coefficients (vectors) and a dictionary. Consider N HSI pixels of size M, such as $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \in \mathbb{R}^{M\times N}$ with SR as

$$\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{Y} \tag{1}$$

where $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n] \in \mathbb{R}^{M\times n}$ denotes the dictionary and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{n\times N}$ represents the sparse coefficient matrix. Each sparse coefficient $\boldsymbol{y} \in \mathbb{R}^n$ in matrix $\boldsymbol{Y}$ is sparse and contains only a few nonzero entries. Each $\boldsymbol{d} \in \mathbb{R}^M$ is called an atom of the dictionary.

Without considering the label information of some training pixels, unsupervised DL aims at training a dictionary, which can reduce the representation error while inducing sparsity in the SR model [30] for all input pixels. To serve the purpose, the classical DL technique considers to minimize an empirical cost function

$$\min_{\boldsymbol{D}} f_N(\boldsymbol{D}) \tag{2}$$

where $f_N(\boldsymbol{D}) \triangleq (1/N)\sum_{i=1}^{N} \ell(\boldsymbol{x}_i, \boldsymbol{D})$. The loss function $\ell(\boldsymbol{x}, \boldsymbol{D})$ is defined as $\ell(\boldsymbol{x}, \boldsymbol{D}) \triangleq \min_{\boldsymbol{y}}(1/2)\|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{y}\|_2^2 + \lambda\psi(\boldsymbol{y})$, where λ is a regularization parameter and $\psi(\cdot)$ is a sparsity inducing regularizer (e.g., well-known ℓ_1 norm [31], [32] and ℓ_0 norm [40], [41]). Problem in (2) can be rewritten as a joint optimization problem as

$$\min_{\boldsymbol{D},\boldsymbol{Y}} \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{y}_i\|_2^2 + \lambda\psi(\boldsymbol{y}_i)\Big). \tag{3}$$

A practical optimization strategy can be found by splitting the problem into two parts, which are alternately solved within an iterative loop. The two parts are: 1) *Sparse Coding*—keeping dictionary $\boldsymbol{D}$ fixed and solving all $\{\boldsymbol{y}_i\}(i = 1, 2, \ldots, N)$ and 2) *Dictionary Update*—keeping $\{\boldsymbol{y}_i\}$ fixed and updating dictionary $\boldsymbol{D}$.

III. Proposed Method

In this section, the motivation of the proposed method is first demonstrated. Then, the proposed method named CODL is introduced.

A. Motivation

Several observed issues in DL and our considerations with respect to those issues are stated as follows.

1) The model of DL in (3) regards each HSI pixel as independent signal and only exploits its spectral signatures to learn the dictionary and obtain sparse coefficients. However, this is not enough for HSIs classification due to the well-known "nonsmooth" characteristic [42], [43] of sparse coding, which means a small variation in the original space might lead to a large difference in the sparse code space. Due to the influence of spectral mixing and noises, the spectral signatures of pixels from the same class could have some small variations. Therefore, although pixels from the same class are similar, they may be represented by very different sparse coefficients. As a result, the discriminative ability of sparse coefficients for classification is degraded. To suppress the intraclass variation, we exploit the contextual characteristics [23] of HSIs that neighboring and similar pixels usually represent the same material and are from the same class. Spectral similarity of neighboring pixels is calculated and used to constrain the differences of pixels' sparse coefficients in the DL model. As a consequence, sparsity patterns of sparse coefficients are corrected in DL by calculating similar sparse coefficients for similar pixels. In this way, intraclass variation can be suppressed, which improves the classification performance.

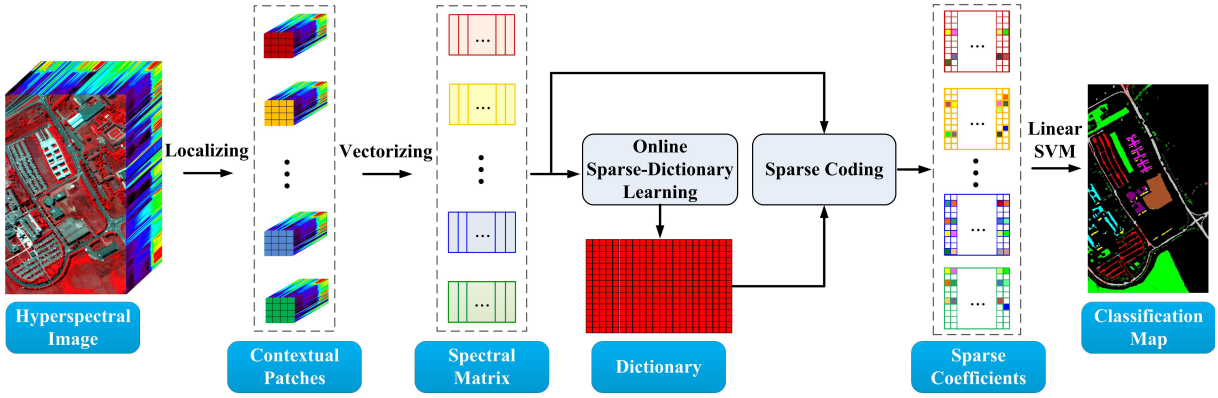


Fig. 1. Flowchart of the proposed method.

2) As shown in Section II, the classical DL mechanism calculates the sparse coding of all pixels in each iteration. In this case, a large number N of pixels can bring high computation cost to DL. Unfortunately, in practical HSI classification tasks, the number N of HSI pixels can be more than 10 000 and thus make DL impractical. We propose a method, which uses an online learning mechanism for DL instead of the classical DL mechanism. The online learning mechanism allow us to input some new samples rather the whole data set for sparse coding and dictionary update in each iteration, which efficiently reduces the computation cost of DL.

B. Contextual Online Dictionary Learning

The flowchart of the proposed CODL method is shown in Fig. 1. The detailed description of CODL method is demonstrated as follows.

1) Dictionary Learning With Contextual Information: It is observed that the pixels in an HSI are usually surrounded by pixels, which are from the same class and similar in terms of spectral characteristics. In this paper, we exploit such contextual information to correct sparsity patterns of sparse coefficients by jointly and sparsely representing a group of neighboring pixels owning similar spectral characteristics. To serve this purpose, a set of nonoverlapping patches [30] with the size of $S \times S$ (called contextual patches) are used to group neighboring pixels. That is, an HSI with N pixels $\boldsymbol{X}$ is divided into K contextual patches $\{\boldsymbol{X}_{\Omega_k}\}_{k=1,\ldots,K}$), where K is determined by the size $S \times S$ of patches (basically equal to $K = N/S^2$). To jointly represent pixels within each contextual patch during DL, the loss function in (2) is extended to

$$\ell(\boldsymbol{X}_\Omega, \boldsymbol{D}) = \min_{\boldsymbol{Y}_\Omega} \frac{1}{2}\|\boldsymbol{X}_\Omega - \boldsymbol{D}\boldsymbol{Y}_\Omega\|_F^2 + \lambda\psi(\boldsymbol{Y}_\Omega) \tag{4}$$

where the $\boldsymbol{Y}_\Omega$ values are sparse coefficient matrix with respect to $\boldsymbol{X}_\Omega$, and prior knowledge about contextual information and sparsity are induced by $\psi(\boldsymbol{Y}_\Omega)$. A traditional way of defining $\psi(\boldsymbol{Y}_\Omega)$ is the *row-sparse* constraint [23], [30], [36], i.e., $\boldsymbol{Y}_\Omega$ has only a few nonzero row. Such constraint regards all pixels in a contextual patch to have high similarity of the same degree, while ignores the situations that contextual patches may locate in detail areas (e.g., near edges) and contain pixels from different classes (i.e., dissimilar pixels). Instead, we introduce a weight to measure the similarity degree between pixels within each contextual patch and then use the weight to induce similar sparse coefficients for similar pixels in $\psi(\cdot)$.

The weight is constructed by normalizing the spectral angle distance (SAD) [44] of pixels with the sigmoid function. Let $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote two pixels in the same patch $\boldsymbol{X}_{\Omega_k}$. The weight $w_{i,j}$ of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is

$$w_{i,j} = \frac{1}{1 + \mathrm{e}^{a(\mathrm{SAD}(\boldsymbol{x}_i, \boldsymbol{x}_j) - b)}} \tag{5}$$

where a and b are positive constants and can be calculated by cross validation. $\mathrm{SAD}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \cos^{-1}(\boldsymbol{x}_i^T \boldsymbol{x}_j / \|\boldsymbol{x}_i\|_2 \cdot \|\boldsymbol{x}_j\|_2)$ is the SAD of two pixels. Note that weight $w_{i,j}$ is a monotone decreasing function with respect to $\mathrm{SAD}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in (5) while smaller SAD means higher spectral similarity. Therefore, the higher value of weight indicates that two pixels are more similar, and vice versa.

To push similar pixels (i.e., pixels with large weight) to have similar and sparse coefficients, the difference of pixels' sparse coefficients combined with weight is used as a penalty term in $\psi(\cdot)$, in addition to a ℓ_1 sparse penalty. As a consequence, the regularizer $\psi(\boldsymbol{Y}_\Omega)$ is carried out as

$$\psi(\boldsymbol{Y}_\Omega) = \|\boldsymbol{Y}_\Omega\|_1 + \frac{\beta}{2\lambda}\sum_{i,j} w_{i,j}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2, \quad \forall \boldsymbol{y}_i,\ \boldsymbol{y}_j \in \boldsymbol{Y}_\Omega \tag{6}$$

where β is a positive tradeoff constant to balance coefficients' sparsity and similarity. Obviously, to minimize $\psi(\boldsymbol{Y}_\Omega)$ in (4), larger weight $w_{i,j}$ constrains coefficients' difference $\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2$ to smaller, which pushes coefficients $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ to be closer. After some algebraic manipulations, $\psi(\boldsymbol{Y}_\Omega)$ can be rewritten as

$$\psi(\boldsymbol{Y}_\Omega) = \|\boldsymbol{Y}_\Omega\|_1 + \frac{\beta}{2\lambda} tr(\boldsymbol{Y}_\Omega \boldsymbol{L} \boldsymbol{Y}_\Omega), \quad \forall \boldsymbol{y}_i,\ \boldsymbol{y}_j \in \boldsymbol{Y}_\Omega \tag{7}$$

where $\boldsymbol{L} = \boldsymbol{C} - \boldsymbol{W}$ is the Laplacian matrix [45]. $\boldsymbol{C}$ is a diagonal matrix, and the ith entry is $c_{ii} = \sum_j w_{ij}$.

2) Online Dictionary Learning: Different with the classical DL mechanism, which inputs all pixels for SR learning (i.e., alternative sparse coding and dictionary update) at each time, online learning mechanism only selects a few of new pixels for SR learning in each iteration. Since most of N pixels are not available in each iteration, the minimization of the empirical cost $f_N(\boldsymbol{D})$ is difficult to be realized. Fortunately, Bottou and Bousquet [33] have pointed out that rather than accurately minimizing the empirical cost, we should focus on the minimization of an expected cost denoted as

$$f(\boldsymbol{D}) \triangleq \mathbb{E}_{\boldsymbol{X}_\Omega}[\ell(\boldsymbol{X}_\Omega, \boldsymbol{D})] \tag{8}$$

where $\mathbb{E}_{\boldsymbol{X}_\Omega}[\cdot]$ means expectation with respect to $\boldsymbol{X}_\Omega$. Let t represent the tth iteration and $\{\boldsymbol{X}_{\Omega_i}\}_{i=1,\ldots,t}$ represent the selected patches in previous iterations. The minimization of expectation in the tth iteration can be approximated by the minimization of the sample average [46]

$$\min_{\boldsymbol{D}} \frac{1}{t}\sum_{i=1}^{t} \ell(\boldsymbol{X}_{\Omega_i}, \boldsymbol{D}). \tag{9}$$

We denote the sample average as $f_t(\boldsymbol{D}) = (1/t)\sum_{i=1}^{t} \ell\ (\boldsymbol{X}_{\Omega_i}, \boldsymbol{D})$. Therefore, as new pixels are input in each iteration, the purpose of DL lies on the minimization of $f_t(\boldsymbol{D})$.

To solve (9) with an online-type method, we introduce a surrogate function $\hat{f}_t(\boldsymbol{D})$ [34] to approximate $f_t(\boldsymbol{D})$. The surrogate function is denoted as

$$\hat{f}_t(\boldsymbol{D}) = \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}\|\boldsymbol{X}_{\Omega_i} - \boldsymbol{D}\hat{\boldsymbol{Y}}_{\Omega_i}\|_F^2 + \lambda\psi(\hat{\boldsymbol{Y}}_{\Omega_i})\right) \tag{10}$$

where the $\{\hat{\boldsymbol{Y}}_{\Omega_i}\}_{i=1,\ldots,t-1}$ values are sparse coefficients obtained in previous iterations, and $\hat{\boldsymbol{Y}}_{\Omega_t}$ is calculated over the dictionary $\boldsymbol{D}_{t-1}$ obtained in previous iteration. That is, in the tth iteration, the sparse coefficient $\hat{\boldsymbol{Y}}_{\Omega_t}$ of $\boldsymbol{X}_{\Omega_t}$ is first calculated by the sparse coding

$$\hat{\boldsymbol{Y}}_{\Omega_t} = \arg\min_{\boldsymbol{Y}_{\Omega_t}} \frac{1}{2}\|\boldsymbol{X}_{\Omega_t} - \boldsymbol{D}_{t-1}\boldsymbol{Y}_{\Omega_t}\|_F^2 + \lambda\psi(\boldsymbol{Y}_{\Omega_t}). \tag{11}$$

Then, $\min_{\boldsymbol{D}} \hat{f}_t(\boldsymbol{D})$ is solved to yield dictionary $\boldsymbol{D}_t$. Such method can be recognized as the "mini-batch"-based extension of online DL in [34].

Next, we show how to solve the sparse coding problem and update $\boldsymbol{D}_t$. Substitute (7) into (11) and yield

$$\hat{\boldsymbol{Y}}_{\Omega_t} = \arg\min_{\boldsymbol{Y}_{\Omega_t}} \frac{1}{2}\|\boldsymbol{X}_{\Omega_t} - \boldsymbol{D}_{t-1}\boldsymbol{Y}_{\Omega_t}\|_F^2 + \|\boldsymbol{Y}_\Omega\|_1 + \frac{\beta}{2}tr(\boldsymbol{Y}_{\Omega_t}\boldsymbol{L}\boldsymbol{Y}_{\Omega_t}). \tag{12}$$

In this paper, we adopt the iterative projective method (IPM) [47] to solve (12). The IPM is able to solve optimization problems in the form of

$$\boldsymbol{Y} = \arg\min_{\boldsymbol{Y}} F(\boldsymbol{Y}) + \lambda J(\boldsymbol{Y}) \tag{13}$$

where F is a convex and smooth function and J is a separable regularizer. Define $F(\boldsymbol{Y}_{\Omega_t})$ as

$$F(\boldsymbol{Y}_{\Omega_t}) = \frac{1}{2}\|\boldsymbol{X}_{\Omega_t} - \boldsymbol{D}_{t-1}\boldsymbol{Y}_{\Omega_t}\|_F^2 + \frac{\beta}{2}tr(\boldsymbol{Y}_{\Omega_t}\boldsymbol{L}\boldsymbol{Y}_{\Omega_t}) \tag{14}$$

which is convex and differentiable. In addition, let $J(\boldsymbol{Y}_{\Omega_t}) = \|\boldsymbol{Y}_{\Omega_t}\|_1$. The sparse coding problem in (12) can be transferred to

$$\hat{\boldsymbol{Y}}_{\Omega_t} = \arg\min_{\boldsymbol{Y}_{\Omega_t}} F(\boldsymbol{Y}_{\Omega_t}) + \lambda J(\boldsymbol{Y}_{\Omega_t}) \tag{15}$$

which can be solved by the IPM. Similar to [45], general sparse coding (i.e., the iterative reweighted methods [48]) can be utilized to initialize coefficients to speed up the convergence. After $\hat{\boldsymbol{Y}}_{\Omega_t}$ is obtained, the dictionary $\boldsymbol{D}_t$ is obtained by solving

$$\begin{aligned}\boldsymbol{D}_t &= \arg\min_{\boldsymbol{D}} \sum_{i=1}^{t}\left(\frac{1}{2}\|\boldsymbol{X}_{\Omega_i} - \boldsymbol{D}\hat{\boldsymbol{Y}}_{\Omega_i}\|_F^2 + \lambda\|\hat{\boldsymbol{Y}}_{\Omega_i}\|_1\right)\\ &= \arg\min_{\boldsymbol{D}} \frac{1}{2}tr(\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{A}_t) - tr(\boldsymbol{D}\boldsymbol{B}_t)\end{aligned} \tag{16}$$

where $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$ are statistic matrixes and record the information of DL in previous iterations. To prevent $\mathbf{D}$ from growing unbounded, atoms of $\mathbf{D}$ are normalized to unit norm as their ℓ_2-norm is larger than one in each iteration of dictionary update. Matrixes $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$ are obtained by

$$\begin{aligned}\boldsymbol{A}_t &= \boldsymbol{A}_{t-1} + \hat{\boldsymbol{Y}}_{\Omega_t}\hat{\boldsymbol{Y}}_{\Omega_t}^T\\ \boldsymbol{B}_t &= \boldsymbol{B}_{t-1} + \boldsymbol{X}_{\Omega_t}\hat{\boldsymbol{Y}}_{\Omega_t}^T.\end{aligned} \tag{17}$$

The block-coordinate descent method [34] is used in this paper to solve (16) and sequently update each atom of the dictionary.

The online DL described earlier selects one contextual patch in each iteration of DL. In practice, considering the large size of HSIs, we improve the convergence speed of the dictionary by selecting p ($p \geq 1$) contextual patches in each iteration. After the iteration is terminated, the unsupervised DL is done and results in the well-learned dictionary. To classify HSIs, sparse coefficients of pixels over the well-learned dictionary are calculated and input into the linear SVM classifier to yield class labels. The proposed CODL method for the classification of HSI is summarized in Algorithm 1.

IV. Experimental Results and Discussions

In this section, we demonstrate both quantitative and visual classification results of the proposed CODL method on three real HSIs. Moreover, the proposed CODL method is compared with some classic or state-of-the-art HSI classification methods in terms of classification performance.

A. Experimental Setup

Three widely used HSIs are used in experiments, including the Pavia University image[1], the Indian Pines image,[1] and Houston image.[2]

[1] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

[2] http://hyperspectral.ee.uh.edu/?page_id=459

Algorithm 1 Proposed CODL Method

Input: 1) The hyperspectral image $\boldsymbol{X}$; 2) Predefined parameters, λ and β; 3) Size of contextual patches S; 4) Number of selected patches for dictionary update p; 5) Labeled training pixels $\boldsymbol{X}_{\text{Train}}$ for classification

Output: Classification result, $\hat{\boldsymbol{Z}}$;

Divide $\boldsymbol{X}$ into K contextual patches, $\{\boldsymbol{X}_{\Omega_k}\}_{k=1,\ldots,K}$;
Generate similarity matrix $\boldsymbol{W} = \{w_{i,j}\}$ by (5);
Initiation: $t = 0$, $p^t = 0$;
for the tth iteration **do**
  Select p contextual patches $\{\boldsymbol{X}_{\Omega_{p^t+1}}, \ldots, \boldsymbol{X}_{\Omega_{p^t+p}}\}$;
  Obtain sparse coefficients $\{\hat{\boldsymbol{Y}}_{\Omega_{p^t+1}}, \ldots, \hat{\boldsymbol{Y}}_{\Omega_{p^t+p}}\}$ by solving (11);
  $t = t + 1$;
  Calculate $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$ according to (17);
  Update dictionary $\boldsymbol{D}_t$ by solving (16) with $\boldsymbol{A}_t$, $\boldsymbol{B}_t$ and $\{\hat{\boldsymbol{Y}}_{\Omega_{p^t+1}}, \ldots, \hat{\boldsymbol{Y}}_{\Omega_{p^t+p}}\}$;
  $p^t = p^{t-1} + p$;
  if $p^t = K$;
    break;
  end
end for
Generate the optimized dictionary, $\hat{\boldsymbol{D}} = \boldsymbol{D}_t$;
Estimate sparse coefficients $\hat{\boldsymbol{Y}}$ based on $\hat{\boldsymbol{D}}$ by solving (11);
Train the linear SVM using sparse coefficients $\hat{\mathbf{Y}}_{\text{Train}}$ with respect to training pixels $\boldsymbol{X}_{\text{Train}}$;
Use the SVM to classify $\hat{\boldsymbol{Y}}$ and obtain the class labels $\hat{\boldsymbol{Z}}$.

1) Pavia University Image: The image is captured by the Reflective Optics System Imaging Spectrometer sensor over the urban area surrounding the University of Pavia, Italy. It consists of 115 spectral bands across the spectral range from 0.43 to 0.86 μm, and each band contains 610×340 pixels with a spatial resolution of 1.3 m. The 12 very noisy channels are removed in the experiment.

2) Indian Pines Image: The image is collected by the Airborne/Visible Infrared Imaging Specrometer senor over the agricultural Indian Pine test site in northwestern Indiana. This HSI consists of 224 spectral bands across the spectral range from 0.4 to 2.5 μm, and each band contains 145×145 pixels with a spatial resolution of 20 m. The 24 water absorption and noise bands are removed in the experiment, i.e., bands no. 1, 33, 97, 107–111, 153–167, and 224.

3) Houston Image: This image is acquired by the Center for Airborne Laser Mapping over the University of Houston campus and the neighboring urban area. It consists of 144 spectral bands across the spectral range from 0.38 to 1.05 μm, and each band contains 349×1300 pixels with a spatial resolution of 2.5 m.

4) Methods for Comparison: Some related methods are used for comparison in experiments, including the superpixel-based classification via multiple kernels (SC-MK) method [19], the LBP and ELM (LBP-ELM) method [13], the extinction profiles (EPs) method [49], the label consistent K-singular value decomposition (KSVD) with statistic features of mean spectra (LC-KSVD$_\mu$) [29], the discriminative KSVD with statistic features of mean spectra (D-KSVD$_\mu$) [50], the fisher discriminant DL with statistic features of mean spectra (FDDL$_\mu$) [47], and the SADL method [30]. The SC-MK, LBP-ELM, and EPs methods are the state-of-the-art spectral–spatial classification methods, which adaptively extract the spectral–spatial features and utilize the advanced classifiers for classification. The LC-KSVD$_\mu$, D-KSVD$_\mu$, FDDL$_\mu$, and SADL are DL-based methods, which simultaneously exploit spectral and spatial information for classification.

TABLE I
PARAMETERS OF THE CODL METHOD USED IN THE EXPERIMENTS WITH RESPECT TO THE THREE HSIs

	$S \times S$	Ratio
Pavia University	18×18	8
Indian Pines	8×8	2
Houston	6×6	2

5) Parameter Setting: The parameters of the SC-MK method, the LBP-ELM method, and the SADL method are set according to the reported parameters in their corresponding references. The EPs method first applies the principal component analysis (PCA) to HSIs and then extract spatial features based on the first of three PCA bands. Note that the parameters of spatial feature extraction are set according to [49]. The LC-KSVD$_\mu$, D-KSVD$_\mu$, and FDDL$_\mu$ methods use the parameters recommended by their references except for the dictionary size and window size (used for mean spectra extraction). The dictionary size is set to 600, 500, and 500 for the Pavia University image, the Indian Pines image, and the Houston image, respectively. Correspondingly, the window size is set in the respective cases to 11, 5, and 5.

For the proposed CODL method, we set the number of selected contextual patches in each iteration to $p = 10$ for the three HSIs. The constants λ and β in the DL model are fixed to $\lambda = 5$ and $\beta = 0.5$. In addition, two important parameters of the proposed CODL method, i.e., the size S of contextual patches, and the ratio used to determine the dictionary size, are empirically chosen based on some related works [23], [30], [35], [40]. The values of S and ratio used in this paper to result in effective and efficient classification for different HSIs are listed in Table I. In experiments, we will analyze the effect of the two parameters (i.e., S and ratio) on the proposed method.

B. Experiments With Pavia University Image

The Pavia University image has nine labeled classes. The false color composite and the label map of the Pavia University image are shown in Fig. 2(a) and (b), respectively. In the experiments, about 9% of labeled pixels are used for training of supervised classification, leaving 91% for test [23], [30]. The detailed number of training and test pixels, as well as the corresponding colors of different classes, is reported in Table II. The color maps of training and test data are demonstrated in Fig. 2(c) and (d), respectively.

The quantitative classification results of the proposed CODL method are reported in Table II. In addition to the classification accuracy for each class, the overall accuracy (OA), the average accuracy (AA), and the Kappa coefficient (κ)) are used for

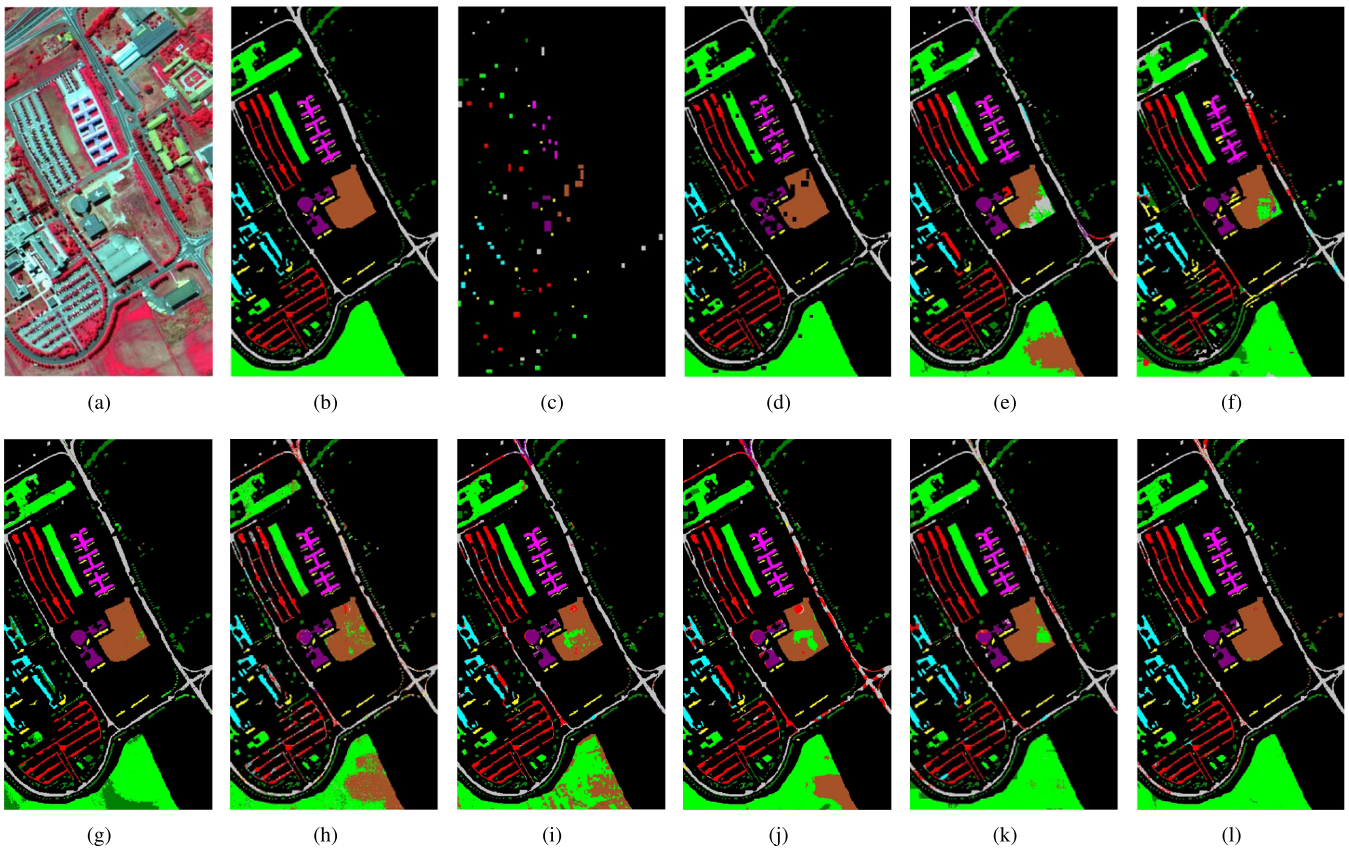


Fig. 2. Pavia University image. (a) False color composite. (b) Reference map. (c) Training data. (d) Test data. (e) SC-MK [19]. (f) LBP-ELM [13]. (g) EPs [49]. (h) LC-KSVD$_\mu$ [29]. (i) D-KSVD$_\mu$ [50]. (j) FDDL$_\mu$ [47]. (k) SADL [30]. (l) CODL.

TABLE II

CLASSIFICATION ACCURACIES (%) OF THE PAVIA UNIVERSITY OBTAINED BY SC-MK [19], LBP-ELM [13], EPS [49], LC-KSVD$_\mu$ [29], D-KSVD$_\mu$ [50], FDDL$_\mu$ [47], SADL [30], AND CODL (THE HIGHEST VALUES FOR EACH CASE ARE GIVEN IN BOLD)

Class	Name	Train	Test	SC-MK	LBP-ELM	EPs	LC-KSVD$_\mu$	D-KSVD$_\mu$	FDDL$_\mu$	SADL	CODL
1	Asphalt	548	6404	88.47	52.78	**97.53**	76.40	81.66	67.42	81.92	87.53
2	Meadows	540	18146	72.9	87.80	81.53	69.64	75.79	82.88	94.43	**97.88**
3	Gravel	392	1815	78.18	68.15	76.47	79.01	75.59	67.11	86.83	**98.24**
4	Trees	524	2912	**97.22**	52.85	90.56	64.01	92.93	93.92	92.51	82.83
5	Metal sheets	265	1113	**100**	89.22	99.73	98.56	98.56	99.91	98.65	**100**
6	Bare soil	532	4572	63.04	78.37	98.75	87.45	86.09	78.06	89.52	**99.28**
7	Bitumen	375	981	85.12	76.66	**100**	90.62	88.99	72.48	79.61	94.50
8	Bricks	514	3364	96.82	71.61	**97.38**	69.74	93.13	88.08	90.55	92.63
9	Shadows	231	795	85.91	55.22	**97.86**	81.64	**97.86**	80.88	**97.86**	92.33
OA	-		-	79.56	75.53	89.07	74.32	81.99	80.60	90.91	**94.76**
AA	-		-	85.30	70.3	93.31	79.67	87.85	81.19	90.21	**93.91**
κ	-		-	0.736	0.675	0.858	0.671	0.767	0.746	0.878	**0.929**

measurement in Table II. The OA is used to measure the ratio of correctly classified pixels in all test pixels, while the AA is the average value of the accuracy with respect to each class. κ is calculated from a confusion matrix and used to measure the agreement of classification. Fig. 2 demonstrates the visual classification maps of various classification methods. Some observations can be made.

According to the maps of training and test pixels in Fig. 2(c) and (d), the training pixels are selected from some small regions, instead of random selection from the reference map. In this case, some feature extraction-based spectral–spatial classification methods (i.e., the SC-MK and LBP-ELM methods), which show outstanding classification performance with randomly selected training pixels, have relatively low classification accuracy, e.g., the OA is 79.56% for the SC-MK method and 75.53% for the LBP-ELM method. The degradation of classification performance is caused by the reason shown in the following. When training pixels are randomly chosen, the selected pixels seem to locate all over the reference map. Each training pixel is surrounded by many test pixels, which usually have the same class label with the training pixel. The information of intraclass correlation

TABLE III

RUNNING TIME (S) OF THE PAVIA UNIVERSITY IMAGE OBTAINED BY LC-KSVD$_\mu$ [29], D-KSVD$_\mu$ [50], FDDL$_\mu$ [47], SADL [30], AND CODL

	LC-KSVD$_\mu$	D-KSVD$_\mu$	FDDL$_\mu$	SADL	CODL
Time (s)	122.8	528.1	549.2	562.0	134.3

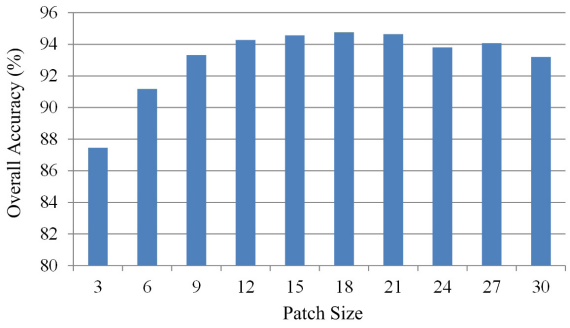


Fig. 3. Different classification accuracies of the CODL method with different sizes of contextual patches, in terms of the Pavia University image.

between each training pixel and its neighboring test pixels is implicitly included into training set and benefits classification. When training pixels locate in some small regions, much less test pixels are adjacent to training pixels. As a consequence, the intraclass correlation between training and test pixels will be weakened and lead to the degradation of the SC-MK and LBP-ELM methods' classification performance.

In addition, three supervised DL-based methods, i.e., the LC-KSVD$_\mu$, D-KSVD$_\mu$, and FDDL$_\mu$ method, learn the discriminative information only from training pixels. With the same reason that less information is included in the training set, the classification accuracies of the three methods are not high. Therefore, these methods seriously depend on the selection of training pixels. In contrast, the unsupervised DL method, i.e., the SADL method, learns the discriminative information from the whole data set without the dependence of training pixels. As a consequence, the SADL method achieves much higher classification accuracy as compared with the supervised DL methods. Similar to the SADL method, the proposed CODL method learns the dictionary over the whole image. Moreover, the CODL method adaptively exploits the contextual information of pixels, which benefits classification. As compared with the SADL method, the CODL method improves the classification accuracy (OA) for about 4%. Experimental results in Table II demonstrate that the CODL method outperforms the state-of-the-art spectral–spatial feature method (EPs method), as well as other comparison methods, in terms of the main metrics (OA, AA, and *kappa*). Therefore, the effectiveness of the proposed CODL method is confirmed.

The running time of the proposed CODL method as well as other DL methods are reported in Table III. All methods are implemented on a laptop, which has an Intel Core CPU 2.50 GHz and 16-GB RAM, in nonparallel implementation. According to Table III, the DL-based methods generally spend

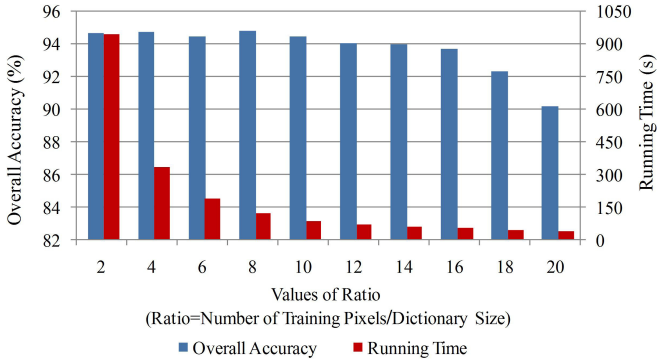


Fig. 4. Classification accuracy and running time of the CODL method over the Pavia University image with different ratios. Ratios are used to determine the size of dictionary with the rule of dividing the number of training pixels by ratios.

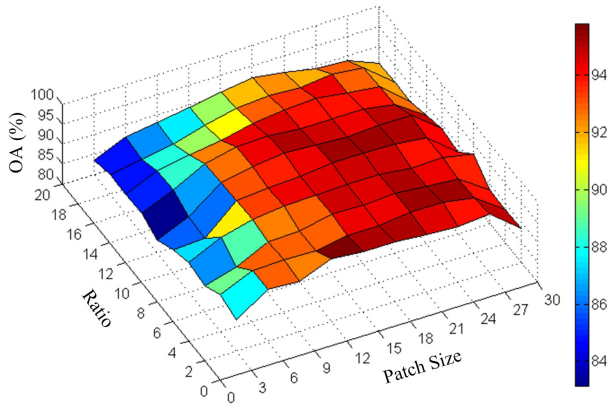


Fig. 5. Classification accuracy (OA) of the proposed CODL method over the Pavia University image, by using different ratios and patch sizes.

more than 100 s for classification. The iterative alternation of sparse coding and dictionary update is time-consuming. Even, the supervised DL methods (i.e., D-KSVD$_\mu$ and FDDL$_\mu$), which have only a few training samples for dictionary construction, own the running time of more than 500 s. Thanks to the exploitation of online learning mechanism, the proposed CODL method is efficient and has only a little longer running time than the LC-KSVD$_\mu$ method, i.e., the difference is less than 12 s. As compared with other DL methods, the CODL method has much less running time and thus shows its efficiency. Next, we will discuss two important parameters in the proposed CODL method, i.e., the size of contextual patch S and the size of dictionary.

The size $S \times S$ of contextual patches is an important parameter for contextual information exploration in this paper. To investigate the influence of this parameter on the proposed method, we conduct experiments by adopting a set of different sizes for the CODL method and collecting corresponding classification accuracy (OA). In this experiment, other parameters are fixed and S is ranged from 3 to 30 with the step size of 3. The obtained classification accuracy in terms of different S values is shown in Fig. 3. As can be observed in Fig. 3, the classification accuracy (OA) of the proposed method gradually increases to a high value as the patch size S increases from 3 to 12. Then, the classification accuracy

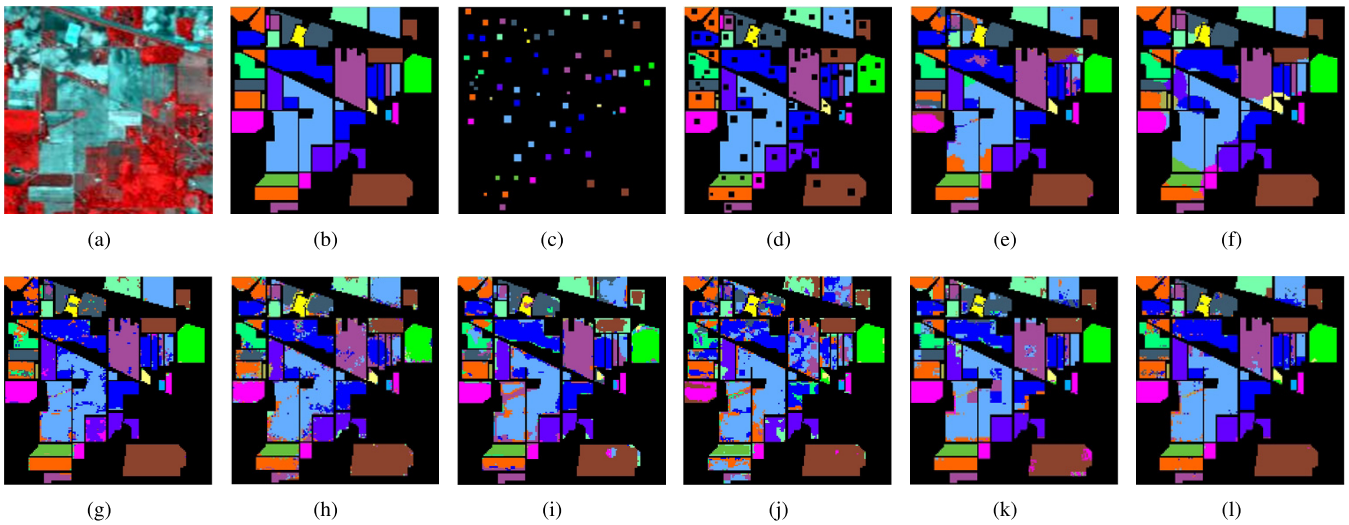


Fig. 6. Classification maps for the Indian Pines Image. (a) False color composite. (b) Reference map. (c) Training samples. (d) Test samples. (e) SC-MK [19]. (f) LBP-ELM [13]. (g) EPs [49]. (h) LC-KSVD$_\mu$ [29]. (i) D-KSVD$_\mu$ [50]. (j) FDDL$_\mu$ [47]. (k) SADL [30]. (l) CODL.

TABLE IV

CLASSIFICATION ACCURACIES (%) OF THE INDIAN PINES IMAGE OBTAINED BY SC-MK [19], LBP-ELM [13], EPS [49], LC-KSVD$_\mu$ [29], D-KSVD$_\mu$ [50], FDDL$_\mu$ [47], SADL [30], AND CODL (THE HIGHEST VALUES FOR EACH CASE ARE GIVEN IN BOLD)

Class	Name	Train	Test	SC-MK	LBP-ELM	EPs	LC-KSVD$_\mu$	D-KSVD$_\mu$	FDDL$_\mu$	SADL	CODL
1	Alfalfa	6	40	**100**	97.5	95.00	85.00	35.00	25.00	72.5	70.00
2	Corn-no till	153	1275	91.92	90.90	89.25	83.61	84.16	59.53	78.90	**92.55**
3	Corn-min till	84	746	77.88	86.60	83.91	75.47	75.07	50.80	87.67	**89.81**
4	Corn	28	209	**96.65**	94.26	66.51	86.12	55.98	62.20	69.86	79.90
5	Grass/pasture-mowed	48	435	70.11	87.82	89.66	88.05	86.21	58.62	**95.40**	94.94
6	Grass/trees	64	666	**100**	82.28	89.49	89.49	90.84	86.19	91.14	98.65
7	Grass/pasture	4	24	95.83	95.83	91.67	79.17	58.33	83.33	**100**	**100**
8	Hay-windrowed	48	430	99.53	99.53	99.30	87.91	82.09	98.14	**100**	98.60
9	Oats	4	16	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
10	Soybean-no till	96	876	**92.24**	90.64	86.64	83.56	89.27	28.88	88.36	91.55
11	Soybean-min till	170	2285	83.28	**90.5**	85.03	80.96	70.94	64.95	81.14	89.06
12	Soybean-clean till	63	530	76.42	78.49	**90.75**	66.04	80.57	38.30	70.19	89.43
13	Wheat	18	187	**100**	87.17	96.26	95.72	90.37	65.78	**100**	99.47
14	Woods	100	1165	92.02	98.37	99.40	96.39	82.83	89.96	91.93	**99.74**
15	Bldg-Grass-Tree-Drives	48	338	96.45	**100**	94.08	88.46	85.21	76.92	92.31	88.46
16	Stone-steel towers	8	85	98.82	92.94	84.71	**100**	72.94	97.65	84.71	92.94
OA	-	-	-	88.30	90.71	89.21	84.38	79.95	64.67	85.62	**92.58**
AA	-	-	-	92.24	92.05	90.1	86,62	77.49	67.89	87.76	**92.19**
κ	-	-	-	0.867	0.894	0.877	0.822	0.773	0.597	0.836	**0.915**

nearly is unchanged when the patch size S is ranged from 12 to 21. After S increases to very high values (e.g., $S = 30$), OA will slightly decline. A too small patch size means that the contextual information is not fully exploited and not sufficient for spectral–spatial classification. Therefore, when $S = 3$ or $S = 6$, the corresponding OA is much lower than the OA for other patch sizes. When S increases to 9, the classification accuracy is significantly improved, demonstrating that the contextual information is very important for the classification. The reason why the classification accuracy decreases for very large patch sizes is explained as follows. The contextual patches with large sizes may contain so much details (e.g., edge areas) that it is difficult to well explore the contextual information in such complex image patches. Above all, the accepted value of S to obtain high and stable classification accuracy can be chosen from a range of $S = 12$ to $S = 21$.

The dictionary size can influence the performance of the dictionary. Large-sized dictionary can be overcomplete [48] for signals and lead to more SR. However, the large-sized dictionary can lead to more complex calculations. On the contrary, small-sized dictionary has less computational cost while may contain insufficient information to sparsely represent some signals (i.e., dictionary is incomplete). To investigate the influence of dictionary size with respect to HSI classification, we conduct experiments by adopting a set of different sizes for dictionary and collecting corresponding classification accuracy (OA) as well as the running time. In this experiment, other parameters are fixed and the dictionary size is changed. We adopt a set of ratios, which range from 2 to 20 with a step size of 2. The dictionary sizes are separately determined by dividing the number of training pixels by those ratios. The classification accuracy and the running time with respect

TABLE V

CLASSIFICATION ACCURACIES (%) OF THE HOUSTON IMAGE OBTAINED BY SC-MK [19], LBP-ELM [13], EPS [49], LC-KSVD$_\mu$ [29], D-KSVD$_\mu$ [50], FDDL$_\mu$ [47], SADL [30], AND CODL (THE HIGHEST VALUES FOR EACH CASE ARE GIVEN IN BOLD)

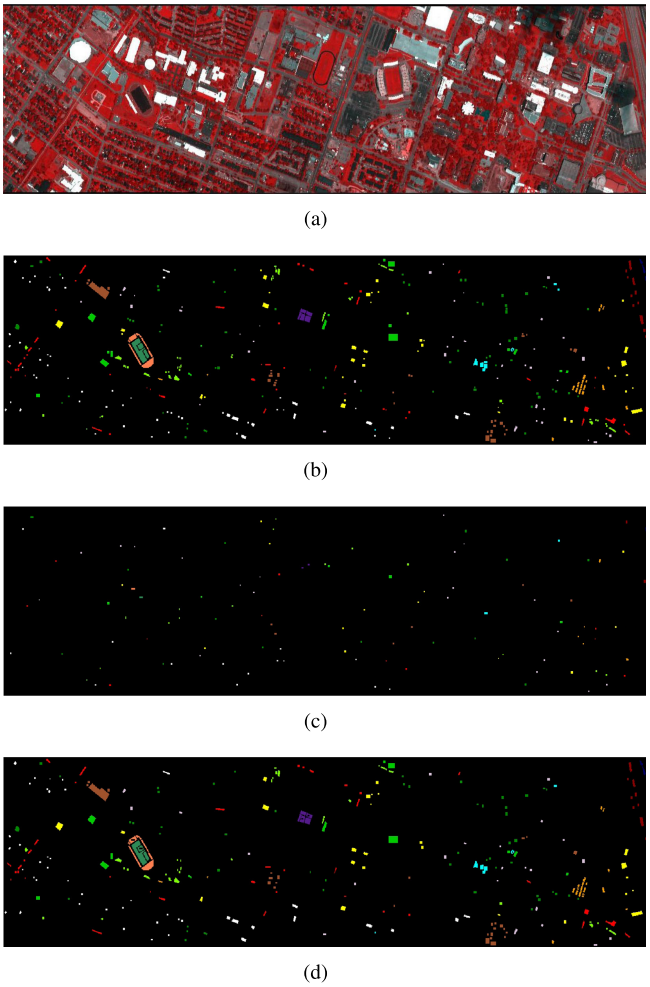
Class	Name	Train	Test	SC-MK	LBP-ELM	EPs	LC-KSVD$_\mu$	D-KSVD$_\mu$	FDDL$_\mu$	SADL	CODL
1	Grass_healthy	100	973	96.71	98.36	97.64	93.73	**100**	**100**	99.38	**100**
2	Grass_stressed	109	701	90.58	97.86	96.01	91.73	**97.29**	96.72	96.72	93.15
3	Grass_synthetic	21	6766	**100**	99.70	**100**	97.19	99.85	90.24	90.53	99.26
4	Tree	164	889	96.4	99.66	98.65	93.81	**100**	97.19	**100**	**100**
5	Soil	73	1169	94.35	**99.49**	74.76	88.45	85.29	86.40	92.81	97.01
6	Water	44	281	61.21	92.17	87.19	77.58	94.66	80.78	91.10	**100**
7	Residential	127	851	79.79	**98.59**	89.07	84.14	86.60	93.65	97.18	95.89
8	Commercial	131	493	75.86	72.62	**79.11**	61.87	68.36	69.78	73.23	69.98
9	Road	84	947	71.91	80.57	78.99	27.67	**85.96**	70.12	69.59	78.88
10	Highway	45	337	55.49	96.14	**100**	83.68	98.81	98.81	**100**	89.61
11	Railway	26	88	28.41	10.23	**100**	**100**	**100**	98.86	61.36	79.55
12	Parking_lot1	128	1105	82.9	64.25	75.11	81.00	79.37	84.8	**92.85**	89.68
13	Parking_lot2	46	403	68.49	25.31	50.37	71.46	74.44	64.02	88.83	**91.81**
14	Tennis_court	20	408	**100**	84.31	**100**	94.85	98.28	89.22	99.02	98.04
15	Running_track	24	636	56.92	97.48	**98.74**	86.95	96.38	95.91	93.24	95.60
OA	-	-	-	83.28	87.32	87.23	81.09	90.21	87.94	91.45	**92.9**
AA	-	-	-	77.27	81.12	88.38	82.27	91.02	87.77	89.72	**91.9**
κ	-	-	-	0.817	0.863	0.861	0.794	0.894	0.869	0.907	**0.923**

to different dictionary sizes are shown in Fig. 4. As can be observed in Fig. 4, the dictionary with small sizes (e.g., ratio is 20) can lead to a low OA with the reason being that the dictionary may be incomplete for some pixels. As the ratio decreases (i.e., dictionary size is increased), the dictionary gradually becomes more complete and leads to higher OA. When the size of the dictionary is large enough (i.e., ratio$\leq$8), the dictionary consists of more atoms and contains enough information to sparsely represent most pixels. In this case, the value of OA only has some small fluctuations. On the other hand, the running time is significantly increased for small ratio (e.g., ratio $=$ 2). To balance the classification performance and computational complexity, we choose ratio $=$ 8 as the optimal value.

Furthermore, we conduct experiments to simultaneously investigate the influence of ratio and patch size S by assigning different values to both parameters. That is, ratio is ranged from 2 to 20 with the step size of 2, and patch size is ranged from 3 to 30 with the step size of 3. The classification accuracies (i.e., OA) of the proposed CODL method using different ratio and patch size are shown in Fig. 5. As can be observed in Fig. 5, high and stable OAs are obtained when ratio is less than 16 and patch size is ranged from 12 to 24. Too small dictionary size and patch size, as well as too large patch size, can result in decreased classification accuracy for the proposed method.

C. Experiments With Indian Pines Image and Houston Image

The Indian Pines image includes 16 labeled classes. The false color composite and the reference map for the Indian Pines image are shown in Fig. 6(a) and (b), respectively. In the experiments, we select around 9% of the labeled pixels for training in supervised classification. The remainder of the labeled pixels is used as test samples. Table IV reports the detailed number of training and test samples, as well as the corresponding colors of different classes. The maps of training and test data are shown in Fig. 6(c) and (d), respectively. The quantitative classification results of the proposed CODL method and other



(a) (b) (c) (d)

Fig. 7. Houston image. (a) False color composite. (b) Reference map. (c) Training data. (d) Test data.

HSIs classification methods are reported in Table IV. The classification map for various classification methods is shown in Fig. 6.

As can be seen in Table IV, the spectral–spatial features-based classification methods, i.e., SC-MK, LBP-ELM methods, and EPs method, show powerful classification performance for the Indian Pines image. Especially, the EPs method has the highest classification accuracies with respect to six classes, i.e., the 3rd, 8th, 10th, 11th, 14th, and 15th classes. The proposed CODL method outperforms these spectral–spatial feature-based methods, as well as other DL methods, in terms of three main indexes, i.e., OA, AA, and κ. Moreover, although other methods have the highest classification accuracies for other classes, the classification accuracies obtained by the proposed CODL method are very close to the highest classification accuracies with respect to some classes, e.g., the 3rd, 5th, 7th, 12th, and 14th classes.

The Houston image has 15 information classes. The false color composite and the label map of this image are shown in Fig. 7(a) and (b), respectively. About 10% of the labeled pixels are used for training and the remainder is used for test. Table V shows the 15 classes and the number of training and test pixels. The visual maps for training and test data are demonstrated in Fig. 7(c) and (d), respectively. The quantitative classification results are reported in Table V. Similar to the Pavia University image, the CODL method obtains the best results in terms of OA, AA, and κ. Besides, the SADL method has comparable performance, that is, the OA of SADL is very close to the OA of CODL (the gap is less than 2%).

V. Conclusion

This paper proposes a DL method named the CODL method for HSIs classification. To efficiently learn the dictionary over a large data set consisted of HSI pixels, the online learning mechanism is introduced to gradually learn the dictionary over a small amount of new input pixels. In addition, the contextual information is adaptively incorporated into online DL by pushing neighboring and similar pixels within nonoverlapping patch to have similar coefficients during the sparsity inducing of DL. The obtained sparse coefficients over the well-learned dictionary are structured, and can be used for SVM classification. Experimental results on several real HSIs show that the proposed method outperforms some classic or state-of-the-art classification methods in terms of classification accuracy.

In this paper, the training of the SVM classifier is not considered in DL. Our future works will focus on simultaneously learning the dictionary as well as the classifier for classification.

References

[1] M. Dalponte, H. O. Orka, T. Gobakken, D. Gianelle, and E. Næsset, "Tree species classification in boreal forests with hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2632–2645, May 2013.

[2] C. M. Gevaert, J. Suomalainen, J. Tang, and L. Kooistra, "Generation of spectral–temporal response surfaces by combining multispectral satellite and hyperspectral UAV imagery for precision agriculture applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3140–3146, Jun. 2015.

[3] L. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Hyperspectral remote sensing image subpixel target detection based on supervised metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4955–4965, Aug. 2014.

[4] Y. Liu, G. Gao, and Y. Gu, "Tensor matched subspace detector for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 1967–1974, Apr. 2017.

[5] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.

[6] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud removal based on sparse representation via multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2998–3006, May 2016.

[7] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[8] F. Ratle, G. Camps-Valls, and J. Weston, "Semisupervised neural networks for efficient hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2271–2282, May 2010.

[9] Y. Yuan, X. Zheng, and X. Lu, "Discovering diverse subset for unsupervised hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 51–64, Jan. 2017.

[10] Y. Zhong and L. Zhang, "An adaptive artificial immune network for supervised classification of multi-/hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 894–909, Mar. 2012.

[11] Y. Gu, T. Liu, X. Jia, J. A. Benediktsson, and J. Chanussot, "Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3235–3247, Jun. 2016.

[12] T. Liu, Y. Gu, X. Jia, J. A. Benediktsson, and J. Chanussot, "Class-specific sparse multiple kernel learning for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7351–7365, Dec. 2016.

[13] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[14] T. Ojala, M. Pietikainen, and T. T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[15] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[16] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[17] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral–spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2565–2574, May 2014.

[18] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

[19] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6663–6674, Dec. 2015.

[20] S. Li, H. Yin, and L. Fang, "Remote sensing image fusion via sparse representations over learned dictionaries," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4779–4789, Sep. 2013.

[21] S. Jia, J. Hu, Y. Xie, L. Shen, X. Jia, and Q. Li, "Gabor cube selection based multitask joint sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3174–3187, Jun. 2016.

[22] E. Zhang, L. Jiao, X. Zhang, H. Liu, and S. Wang, "Class-level joint sparse representation for multifeature-based hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4160–4177, Sep. 2016.

[23] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[24] J. Zou, W. Li, and Q. Du, "Sparse representation-based nearest neighbor classifiers for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2418–2422, Dec. 2015.
[25] W. Li, Q. Du, F. Zhang, and W. Hu, "Hyperspectral image classification by fusing collaborative and sparse representations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4178–4187, Sep. 2016.
[26] S. Li, T. Lu, L. Fang, X. Jia, and J. A. Benediktsson, "Probabilistic fusion of pixel-level and superpixel-level hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7416–7430, Dec. 2016.
[27] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral–spatial classification of hyperspectral images with a superpixel-based discriminative sparse model," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4186–4201, Aug. 2015.
[28] Z. Wang, N. M. Nasrabadi, and T. S. Huang, "Spatial–spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4808–4822, Aug. 2014.
[29] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.
[30] A. Soltani-Farani, H. R. Rabiee, and S. A. Hosseini, "Spatial-aware dictionary learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 527–541, Jan. 2015.
[31] A. S. Charles, B. A. Olshausen, and C. J. Rozell, "Learning sparse codes for hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 963–978, Sep. 2011.
[32] A. Castrodad, Z. Xing, J. B. Greer, E. Bosch, L. Carin, and G. Sapiro, "Learning discriminative sparse representations for modeling, source separation, and mapping of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4263–4281, Nov. 2011.
[33] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2007, pp. 161–168.
[34] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.
[35] H. Zhang, J. Li, Y. Huang, and L. Zhang, "A nonlocal weighted joint sparse representation classification method for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2056–2065, Jun. 2013.
[36] W. Fu, S. Li, L. Fang, X. Kang, and J. A. Benediktsson, "Hyperspectral image classification via shape-adaptive joint sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 2, pp. 556–567, Feb. 2016.
[37] M. Xiong, Q. Ran, W. Li, J. Zou, and Q. Du, "Hyperspectral image classification using weighted joint collaborative representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1209–1213, Jun. 2015.
[38] W. Li, J. Liu, and Q. Du, "Sparse and low-rank graph for discriminant analysis of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4094–4105, Jul. 2016.
[39] T. Lu, S. Li, L. Fang, Y. Ma, and J. A. Benediktsson, "Spectral–spatial adaptive sparse representation for hyperspectral image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 373–385, Jan. 2016.
[40] 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.
[41] A. M. Tillmann, "On the computational intractability of exact and approximate dictionary learning," *IEEE Signal Process. Lett.*, vol. 22, no. 1, pp. 45–49, Jan. 2015.
[42] 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. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3555–3561.
[43] Y. He, K. Kavukcuoglu, Y. Wang, A. Szlam, and Y. Qi, "Unsupervised feature learning by deep sparse coding," in *Proc. SIAM Int. Conf. Data Mining*, 2014, pp. 902–910.
[44] W. Fu, S. Li, L. Fang, and J. A. Benediktsson, "Adaptive spectral–spatial compression of hyperspectral image with sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 671–682, Feb. 2017.
[45] S. Gao, I. W. H. Tsang, and L. T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.
[46] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *Math. Program.*, vol. 157, no. 2, pp. 515–545, Jun. 2016.
[47] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 543–550.
[48] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, Apr. 2010.
[49] P. Ghamisi, R. Souza, J. A. Benediktsson, X. X. Zhu, L. Rittner, and R. A. Lotufo, "Extinction profiles for the classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5631–5645, Oct. 2016.
[50] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2691–2698.

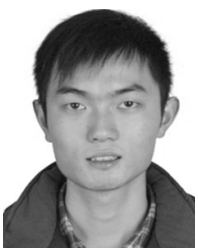
Wei Fu (S'14) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, in 2012, where he is currently pursuing the Ph.D. degree in electrical engineering.

From 2016 to 2017, he was a Visiting Ph.D. Student with the Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavik, Iceland, supported by the China Scholarship Council. His research interests include hyperspectral image processing, compressive sensing, and sparse representation.

Shutao Li (M'07–SM'15) received the B.S., M.S., and Ph.D. degrees from Hunan University, Changsha, China, in 1995, 1997, and 2001, respectively.

From 2002 to 2003, he was a Post-Doctoral Fellow with the Royal Holloway College, University of London, London, U.K., with Prof. J. Shawe-Taylor. In 2005, he visited the Department of Computer Science, The Hong Kong University of Science and Technology, Hong Kong, as a Visiting Professor. He joined the College of Electrical and Information Engineering, Hunan University, in 2001, where he is currently a Full Professor. He was a Research Associate with the Department of Computer Science, The Hong Kong University of Science and Technology, in 2011. He has authored or co-authored over 160 refereed papers. His research interests include compressive sensing, sparse representation, image processing, and pattern recognition.

Dr. Li was a recipient of two 2nd-Grade National Awards at the Science and Technology Progress of China in 2004 and 2006. He is currently an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, and a member of the editorial board of the *Information Fusion* and *Sensing and Imaging*.



Leyuan Fang (S'10–M'14–SM'17) received the B.S. and Ph.D. degrees from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2008 and 2015, respectively.

From 2011 to 2012, he was a Visiting Ph.D. Student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council. Since 2017, he has been an Associate Professor with the College of Electrical and Information Engineering, Hunan University. His research interests include sparse representation and multiresolution analysis in remote sensing and medical image processing.

Dr. Fang received the Scholarship Award for Excellent Doctoral Student granted by the Chinese Ministry of Education in 2011.

Jón Atli Benediktsson (S'84–M'90–SM'99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

From 2009 to 2015, he was the Pro Rector of Science and Academic Affairs and a Professor of electrical and computer engineering with the University of Iceland. In 2015, he became the Rector of the University of Iceland. He is also the Co-Founder of a biomedical startup company Oxymap. He has published extensively in his fields of interest. His research interests include remote sensing, image analysis, pattern recognition, biomedical analysis of signals, and signal processing.

Dr. Benediktsson is a member of the Association of Chartered Engineers at Iceland (VFI), the Societas Scinetiarum Islandica, and the Tau Beta Pi. He is a fellow of the International Society for Optics and Photonics. He was a recipient of the Stevan J. Kristof Award from Purdue University in 1991 as an Outstanding Graduate Student in Remote Sensing. He was also a recipient of the Icelandic Research Councils Outstanding Young Researcher Award in 1997, the IEEE Third Millennium Medal in 2000, the Yearly Research Award from the Engineering Research Institute at the University of Iceland in 2006, the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society in 2007, and the IEEE/VFI Electrical Engineer of the Year Award in 2013. He was a co-recipient of the University of Iceland's Technology Innovation Award in 2004, the 2012 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Paper Award, the IEEE Geoscience and Remote Sensing Society (GRSS) Highest Impact Paper Award in 2013, and the *International Journal of Image and Data Fusion* Best Paper Award in 2014. He was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2007 to 2010. He was the 2011–2012 President of the IEEE GRSS and has been on the GRSS Administrative Committee since 2000. He was the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008 and has been an Associate Editor of the IEEE TGRS since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and the IEEE ACCESS since 2013. He serves on the Editorial Board of the IEEE PROCEEDINGS and the International Editorial Board of the *International Journal of Image and Data Fusion*.