Neurocomputing 226 (2017) 90–100

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification

Xiangyong Cao, Lin Xu, Deyu Meng*, Qian Zhao, Zongben Xu

School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, PR China

ARTICLE INFO

Communicated by Yue Gao

MSC:
00-01
99-00

Keywords:
Hyperspectral image classification
3-dimensional discrete wavelet transform
Support vector machine
Segmentation

ABSTRACT

Hyperspectral image (HSI) classification is one of the fundamental tasks in HSI analysis. Recently, many approaches have been extensively studied to improve the classification performance, among which integrating the spatial information underlying HSIs is a simple yet effective way. However, most of the current approaches haven't fully exploited the spatial information prior. They usually consider this prior either in the step of extracting spatial feature before classification or in the step of post-processing label map after classification, while don't integratively employ the prior in both steps, which thus leaves a room for further enhancing their performance. In this paper, we propose a novel spectral-spatial HSI classification method, which fully utilizes the spatial information in both steps. Firstly, the spatial feature is extracted by applying the 3-dimensional discrete wavelet transform (3D-DWT). Secondly, the local spatial correlation of neighboring pixels is modeled using Markov random field (MRF) based on the probabilistic classification map obtained by applying probabilistic support vector machine (SVM) to the extracted 3D-DWT feature in the first step, and then a maximum a posterior (MAP) classification problem can be formulated in a Bayesian perspective. Finally, α-Expansion min-cut-based optimization algorithm is adopted to solve this MAP problem efficiently. Experimental results on two benchmark HSIs show that the proposed method achieves a significant performance gain beyond state-of-the-art methods.

1. Introduction

Hyperspectral imaging has opened up new opportunities for analyzing a variety of materials due to the rich information on spectral and spatial distributions of the distinct materials in hyperspectral imagery, such as land-use or land-cover mapping, forest inventory, and urban-area monitoring [1]. Many hyperspectral applications can be essentially converted to a classification task, which aims to classify the image pixels of a hyperspectral image into multiple categories. Multiple state-of-the-art classification techniques have been attempted for this task and achieved good performance in certain applications [2–4].

However, these methods still tend to encounter some problems in practical scenarios. First, the available labeled training samples in HSI classification are typically limited because of the expensive image labeling cost [5], which leads to the high-dimension while low-sample-size classification issue. Second, despite the high spectral resolution of HSI, identical material may have quite different spectral signatures, whereas different materials may share similar spectral signatures [6]. The aforementioned problems, coupled with other difficulties, such as embedded noises from the sensors and environment, incline to further decrease the classification accuracy.

Various classification approaches have been investigated to address these problems. A main approach is to discover the essential discriminant features that are beneficial to classification while reducing the noise embedded in HSIs that impairs the classification performance. As to exploiting discriminant features, it has been pointed out in [7] that spatial information is more crucial than the spectral signatures in HSI classification. Therefore, a pixel-wise classification method following a spatial-filtering preprocessing step becomes a simple yet effective way to implement this technique [4,8]. As to spatial-filtering strategy, square patch is a representative method [9–11], which groups the neighboring pixels by square windows firstly and then extracts features based on the local window using other subspace learning techniques, such as low-rank matrix factorization [12–17], dictionary learning [4,18] and subspace clustering [19]. Compared with the original spectral signatures, the filtered features extracted by square patch method have less intra-class variability and higher spatial smoothness, with reduced noise in some sense.

* Corresponding author.
E-mail addresses: caoxiangyong45@gmail.com (X. Cao), xulinshadow@gmail.com (L. Xu), dymeng@mail.xjtu.edu.cn (D. Meng), timmy.zhaoqian@gmail.com (Q. Zhao), zbxu@mail.xjtu.edu.cn (Z. Xu).

http://dx.doi.org/10.1016/j.neucom.2016.11.034
Received 1 July 2016; Received in revised form 1 November 2016; Accepted 18 November 2016
Available online 21 November 2016
0925-2312/

Except spatial-filtering method, another popular approach to exploit the spatial information is spectral-spatial method, which combines the spectral and spatial information into a classifier. Unlike the pixel-wise classification approach that does not take spatial structure information into consideration, this approach incorporates the local consistency of the neighboring pixel labels. Such a methodology has been proposed to unify the pixel-wise classification with a segmentation map [20]. Besides, Markov random field (MRF) is also a popular technique for this spectral-spatial method, which can be applied to incorporate the local correlation of neighboring pixels into the classification step. Then, the classification task can be formulated into a maximum a posterior (MAP) problem [21,22]. Finally, the graph-based segmentation algorithm [23] can be used to solve this problem efficiently. Compared with the methods that don't utilize the spatial contextual information of labels, the MRF-based HSI classification methods have been validated to achieve higher classification accuracy [21,22].

As to the above two approaches of improving the classification performance, most of the current methods usually utilize only one of them, namely, either using the spatial-filtering approach in feature extraction or spectral-spatial approach in embedding the MRF prior into the classification step. These methods still may not help us obtain the best performance due to the lack of fully utilizing the spatial information. In order to more comprehensively exploit the spatial information, a natural idea is to incorporate both of the two methodologies into a unique framework to further prompt the capability of state-of-the-arts for this HSI classification task.

Specifically, in this paper, a novel approach is proposed for supervised HSI classification. The key idea is to unify the spatial-filtering technique and the spatial smoothness (MRF) prior of labels into one framework. Firstly, a spatial-filtering method is used to produce spectral-spatial features. In our work, the 3-dimensional discrete wavelet transform (3D-DWT) [24] is adopted to generate spectral-spatial features, which have been validated to be more discriminative than the original spectral signature [7]. Secondly, the local correlation of neighboring pixels should also be introduced in order to fully exploit the spatial information. In our paper, the Markov random field (MRF), which assumes that adjacent pixels are more likely to belong to the same class, is utilized to model this local correlation. To our best knowledge, this is the first work that spatial feature extraction using 3D-DWT and spatial post-processing using MRF are taken into consideration simultaneously. Besides, probability support vector machine (SVM) is firstly utilized on the spatial 3D-DWT feature. Our extensive experimental results substantiate that such amelioration is insightful to this issue and can always guarantee a considerable improvement beyond the state-of-the-art methods in the scenarios both with relatively less labeled samples and with noisy input training samples. The proposed approach is thus expected to further prompt the frontier of this line of study.

The rest of the paper is organized as follows. In Section 2, related work regarding hyperspectral image (HSI) classification is introduced. In Section 3, the 3-dimensional discrete wavelet transform (3D-DWT) is briefly reviewed. In Section 4, the whole classification method is described. In Section 5, experimental results on two benchmark HSIs are reported. Finally, conclusions are drawn in Section 6. Throughout the paper, we denote scalars, vectors, matrices and tensors as the non-bold, bold lower case, bold upper case and curlicue letters, respectively.

2. Related work

The past two decades has witnessed prosperous developments in the field of hyperspectral image (HSI) classification. Numerous extensions along this line of research can be roughly classified into three categories: support vector machine (SVM)-based methods, sparse representation classifier (SRC)-based methods and multinomial logistic regression (MLR)-based methods.

In HSI classification, support vector machine (SVM) is a state-of-the-art approach that has shown impressive performance in high dimensional scenario [2]. The effectiveness of SVM largely depends on the choice of kernel functions, among which radial basis function (RBF) is the most widely used one. However, SVM [2] is only a pixel-wise classification method and ignores the correlations among distinct pixels in the image, and thus always results in unsatisfying classification performance. To improve the performance of SVM, multiple improved versions have been proposed, including SVM with composite kernels (SVM-CK) [25], which combines both spectral and spatial information in kernels, and multiple kernel learning (MKL) [26], which enhances the flexibility of kernels in machine learning. Except the popular SVM classifier, sparse representation classifier (SRC) has also been widely used in HSI classification [4,27]. Specifically, SRC method is based on the observation that hyperspectral pixels belonging to the same class approximately lie in the same low-dimensional subspace, and then an unknown test sample can be sparsely represented by the combination of a few training samples from the entire dictionary, while the corresponding sparse representation vector encodes the class information implicitly. Many improved versions based on SRC method have also been conducted to discover the inherent structure of adjacent pixels, including joint sparsity model (JSM) [28], which assumes small neighborhood pixels share a common sparsity support, Laplacian regularized Lasso [4], which introduces another weighting matrix to characterize the similarity among neighboring pixels based on JSM, and collaborative group Lasso (CGL) [11], which assumes that the representation matrix has a group-wise sparsity pattern and further enforce sparsity within each group. Except the two previous classifiers, multinomial logistic regression (MLR)-based classifier [29] has also been adopted in HSI classification. It aims to maximize the posterior class distributions for each sample and seems more suitable to the multi-classification task of HSI. Many studies on applying MLR to the HSI classification have obtained promising results [7]. Other utilized methods for HSI classification have also been studied, such as convolutional neural network (CNN) [30], extreme learning machine (ELM) [31], boosting [32] and semi-supervised method [33].

In summary, the performance of HSI classification can be significantly improved by integrating spatial information. More concretely, the methods for utilizing the spatial information of the hyperspectral image can be roughly divided into two main categories, namely spatial-filtering method [4,11] and spectral-spatial method [12,22,21]. For the spatial-filtering strategy, square patch is a representative approach [9–11]. This method can be used to group the neighboring pixels and then these neighboring pixels are applied in SRC and low-rank-based methods. For example, spatial correlation between neighboring pixels is utilized in [28], while square patches are taken as contextual groups in [10]. As to spectral-spatial methods, Markov random field (MRF) model is a commonly used strategy, which can be adopted to incorporate spatial information into the classification step by adding a smoothness prior term of labels on a probabilistic discriminative classification function [12,21,22]. Additionally, conditional random field (CRF) [34] is also an alternative to formulate the spectral-spatial classification model.

Although most of these aforementioned approaches achieve good performance in some scenarios, they still lack of fully utilizing the spatial information, which is a main drawback for these methods. Specifically, most of the current methods utilize the spatial knowledge either in the pre-processing feature extraction stage or in the smoothness post-processing the recovered labels, while not integratively take both useful information into a unique framework, such as SOMP [4], CGL [11] and MKL [26], which only utilize the square patch method to extract spatial information, SVM-GC [22,35] and MLRsubMLL [21], which just utilize the MRF to model the smoothness prior of labels. In order to fully exploit the spatial information, a natural idea is to incorporate both of the two techniques into the classification process. By fully discovering the spatial information, the proposed method can

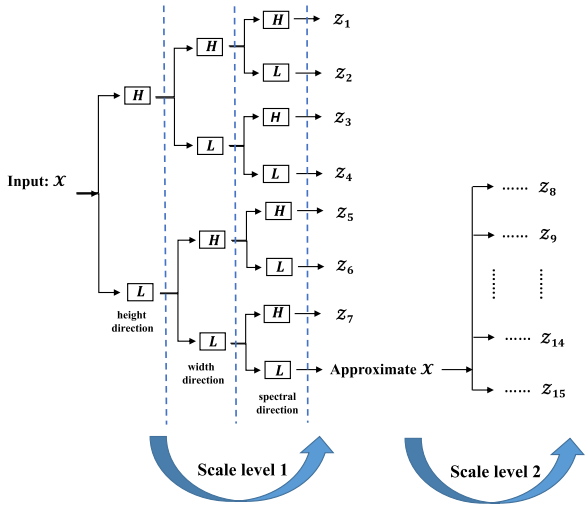


Fig. 1. Flowchart of 3-dimensional DWT.

achieve better classification accuracy in few labeled training samples scenario and noisy scenario, as substantiated by our experiments presented in Section 5.

3. 3-Dimensional discrete wavelet transform

In this section, we first define some notations used throughout this paper, and then introduce the 3-dimensional discrete wavelet transform (3D-DWT) to extract features from HSIs.

Let us denote the given HSI data as $\mathcal{X} \in R^{H\times W\times B}$, where H and W are the height and width of the spatial dimensions, respectively, and B is the number of spectral bands. Assume that the observed training samples (spectral vectors) are denoted as $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in R^{B\times n}$, $n \le HW$, $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in T^n$ is the labels for each observed samples and $T = (1, 2, \ldots, K)$ is the set of class labels.

Wavelet transform (WT) [36] is a popular mathematical tool for time-frequency analysis and has gained widespread acceptance in signal processing and image compression. It is defined as

$$(W_{\phi f})(a, b) = \langle f(x), \phi_{a,b}(x)\rangle = \int f(x)\phi_{a,b}(x)dx, \tag{1}$$

where $\phi_{a,b}(x)$ is wavelet basis function, which is obtained from a single prototype wavelet $\phi(x)$ called mother wavelet by dilations and shifting:

$$\phi_{a,b}(x) = \frac{1}{\sqrt{a}}\phi\left(\frac{x-b}{a}\right), \tag{2}$$

where a is the scaling parameter and b is the shifting parameter. If a and b are discrete values, the discrete wavelet transform (DWT) is given by

$$(W^{\phi}_{m,n})(f) = \langle f(x), \phi_{m,n}(x)\rangle = \int f(x)\phi_{m,n}(x)dx, \tag{3}$$

where $\phi_{m,n}(x) = a_0^{-m/2}\phi(x - nb_0a_0^m/a_0^m)$, a_0 and b_0 are dyadic scale parameter and shifting parameter, respectively. From the perspective of multi-scale analysis, the function $f(x)$ can be recovered from a linear combination of wavelet and scaling functions $\phi(x)$ and $\psi(x)$. Specifically, we can approximate a discrete signal $f[n]$ in $l^2(\mathbf{Z})$[1] by

$$f[n] = \frac{1}{\sqrt{M}}\sum_k C_\psi[j_0, k]\psi_{j_0,k}[n] + \frac{1}{\sqrt{M}}\sum_{j=j_0}^{\infty}\sum_k D_\phi[j, k]\phi_{j,k}[n], \tag{4}$$

where $f[n]$, $\psi_{j_0,k}[n]$ and $\phi_{j,k}[n]$ are discrete functions defined on $[0, M-1]$, containing totally M points. Since the sets $\{\psi_{j_0,k}[n]\}_{k\in\mathbf{Z}}$ and $\{\phi_{j,k}[n]\}_{(j,k)\in\mathbf{Z}^2, j\ge j_0}$ are orthogonal to each other, we can simply take the inner product to obtain the wavelet coefficients by

$$C_\psi[j_0, k] = \frac{1}{\sqrt{M}}\sum_n f[n]\psi_{j_0,k}[n], \tag{5}$$

$$D_\phi[j, k] = \frac{1}{\sqrt{M}}\sum_n f[n]\phi_{j,k}[n]. \tag{6}$$

In our work, we apply the 3-dimensional discrete wavelet transform (3D-DWT) to the hyperspectral cube following [7], which can encode the spatial information into different scales, frequencies and orientations. It should also be noted that 3D-DWT can be achieved by three 1D-DWTs. In our proposed method, the Haar wavelet is used. In practice, wavelet and scaling functions $\psi(x)$ and $\phi(x)$ are represented by the filter bank (L,H) given by the low-pass and high-pass filer coefficients $l[k]$ and $h[k]$, respectively. Specifically, for the Haar wavelet, $l[k] = (1/\sqrt{2}, 1/\sqrt{2})$, and $h[k] = (-1/\sqrt{2}, 1/\sqrt{2})$. At each scale level, the convolution products with all combinations of high-pass and low-pass filters in three dimensions produce eight different filtered hyperspectral cubes. Then, the hyperspectral cube filtered by the low-pass filter in each of the three dimensions is further convolved in the next scale level. In our work, the hyperspectral data cube is only decomposed into two levels, and thus 15 subcubes $\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_{15}$ are generated in total. It should be noted that we have left out the down-sampling step in the standard 1D-DWT and thus each subcube has the same size with the original HSI cube.

The wavelet coefficients of each pixel at position (i,j) in all subcubes can be directly concatenated to form its feature vector

$$\mathbf{x}_{i,j} = (\mathcal{Z}_1(i, j, \cdot), \mathcal{Z}_2(i, j, \cdot), \ldots, \mathcal{Z}_{15}(i, j, \cdot)). \tag{7}$$

Then, to utilize the spatial smoothness, we apply a mean filter to the absolute values of wavelets coefficients

$$\widehat{\mathcal{Z}}_n(i, j, \cdot) = \frac{1}{9}\sum_{a=i-1}^{i+1}\sum_{b=j-1}^{j+1}|\mathcal{Z}_n(a, b, \cdot)|, \; n = 1, 2, \ldots, 15. \tag{8}$$

Let $\widehat{\mathcal{Z}} \in R^{H\times W\times 15B}$ to be the final concatenated cube and the 3D-DWT-based feature vector of pixel (i,j) in $\widehat{\mathcal{Z}}$ is given by

$$\widehat{\mathcal{Z}}(i, j, \cdot) = (\widehat{\mathcal{Z}}_1(i, j, \cdot), \widehat{\mathcal{Z}}_2(i, j, \cdot), \ldots, \widehat{\mathcal{Z}}_{15}(i, j, \cdot)). \tag{9}$$

The main steps of generating features for each pixel using 3D-DWT are shown in Fig. 1.

4. Spectral-spatial classification method using 3D-DWT feature

The goal of image classification is to estimate $\mathbf{y}$ when $\mathbf{x}$ is observed. In the Bayesian framework, the estimation of $\mathbf{y}$ for observations of $\mathbf{x}$ can be conducted by maximizing the posterior distribution [21,37]

$$\mathbb{P}(\mathbf{y}|\mathbf{x}) \propto \mathbb{P}(\mathbf{x}|\mathbf{y})\mathbb{P}(\mathbf{y}), \tag{10}$$

where $\mathbb{P}(\mathbf{x}|\mathbf{y})$ is the likelihood function, namely the probability of the observed feature $\mathbf{x}$ given labels $\mathbf{y}$, and $\mathbb{P}(\mathbf{y})$ denotes the prior knowledge on labels $\mathbf{y}$.

For simplicity, we assume that the features given the class labels are conditionally independent. Therefore, the likelihood function $\mathbb{P}(\mathbf{x}|\mathbf{y})$ can be written as

$$\mathbb{P}(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{n}\mathbb{P}(\mathbf{x}_i|y_i). \tag{11}$$

Then, the posterior $\mathbb{P}(\mathbf{y}|\mathbf{x})$ can be equivalently converted into

$$\mathbb{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathbb{P}(\mathbf{x})}\prod_{i=1}^{n}\mathbb{P}(\mathbf{x}_i|y_i)\mathbb{P}(\mathbf{y}), \qquad = C(\mathbf{x})\prod_{i=1}^{n}\frac{\mathbb{P}(y_i|\mathbf{x}_i)}{\mathbb{P}(y_i)}\mathbb{P}(\mathbf{y}), \tag{12}$$

where $C(\mathbf{x}) = \frac{\prod_{i=1}^{n}\mathbb{P}(\mathbf{x}_i)}{\mathbb{P}(\mathbf{x})}$ is a factor that has no relationship with labels $\mathbf{y}$. In our proposed method, we assume that each class has equal probability, namely $\mathbb{P}(y_i = k) = \frac{1}{K}$, $\forall\, k \in \{1, 2, \ldots, K\}$. Besides, any

[1] $l^2(\mathbf{Z}) = \{f[n] | \sum_{n=-\infty}^{\infty}|f[n]|^2 < \infty\}$.

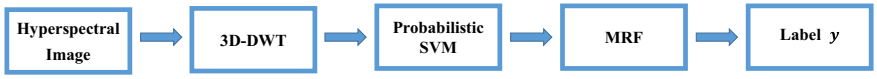


Fig. 2. Flowchart of SVM-3DG.

other distribution assumption of $\mathbb{P}(y_i)$ can also be given. Therefore, by maximizing the posterior distribution, the classification result can be given by

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\left\{\sum_{i=1}^{n} \log \mathbb{P}(y_i|\mathbf{x}_i) + \log \mathbb{P}(\mathbf{y})\right\}. \tag{13}$$

In our approach, the first term $\mathbb{P}(y_i|\mathbf{x}_i)$ (i.e. the class probabilities for each pixel), can be modeled using probabilistic SVM [38,39], which has excellent performance and has been widely adopted to analyze hyperspectral data [40,22,41,12]. Specifically, we use the ith pixel of the stacked-cube $\widehat{\mathcal{Z}}$ obtained by 3D-DWT as the input feature to SVM instead of the ith pixel vector $\mathbf{x}_i$ in the original cube $\mathcal{X}$, which is different from the methods in [22,42]. We denote the ith pixel vector of the stacked-cube $\widehat{\mathcal{Z}}$ as $\mathbf{z}_i$ and replace $\mathbb{P}(y_i|\mathbf{x}_i)$ with $\mathbb{P}(y_i|\mathbf{z}_i)$. Then, the maximum a posterior (MAP) problem (13) can be rewritten as

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\left\{\sum_{i=1}^{n} \log \mathbb{P}(y_i|\mathbf{z}_i) + \log \mathbb{P}(\mathbf{y})\right\}. \tag{14}$$

For the second term $\mathbb{P}(\mathbf{y})$ (i.e. the prior knowledge of the labels), we can utilize the local correlation of labels, which means that it is very likely that the spatially adjacent pixels belong to the same class. Thus $\mathbb{P}(\mathbf{y})$ can be modeled with a Markov random field (MRF) prior which is formulated as

$$\mathbb{P}(\mathbf{y}) = C \exp\left\{-\sum_{(i,j)\in\mathcal{N}_e} W(y_i, y_j)\right\}, \tag{15}$$

where C is a normalization parameter and $W(y_i, y_j)$ is an interaction term expressing spatial coherency between neighboring pixels i and j and $\mathcal{N}_e$ represents the set of neighboring pixels (8-neighborhood is used in our work). To compute the spatial interaction term $W(y_i, y_j)$, a common choice is the Potts model [43] which is defined as

$$W(y_i, y_j) = \beta(1 - \delta(y_i, y_j)), \tag{16}$$

where $\delta(\cdot)$ is the Kronecker function ($\delta(a, b) = 1$ for a=b and $\delta(a, b) = 0$ otherwise), and β is a nonnegative constant parameter that controls the level of spatial smoothness. However, this interaction term tends to deteriorate classification results at the edges between land-cover classes. Therefore, we adopt the new interaction term proposed in [22], that is

$$W(y_i, y_j) = \beta(1 - \delta(y_i, y_j))\exp\{-d(\mathbf{z}_i, \mathbf{z}_j)\}, \tag{17}$$

where $d(\mathbf{z}_i, \mathbf{z}_j)$ measures the dissimilarity between $\mathbf{z}_i$ and $\mathbf{z}_j$. Eq. (17) actually incorporates the information associated with class edge within MRF framework. It tends to make adjacent pixels have the same label, which means it encourages the classification boundary to align with strong edges. Specifically, for neighboring pixels across a strong edge, W_{ij} is larger, which means that y_i and y_j can take different labels after model optimization. But for neighboring pixels within image flat region, W_{ij} is small, which enforces that the neighboring pixels should have the same class label after model optimization. In our work, the dissimilarity measure $d(\mathbf{z}_i, \mathbf{z}_j)$ is defined as

$$d(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{15B}\sum_{b=1}^{15B}\left(q_b(\mathbf{z}_i)\log\frac{q_b(\mathbf{z}_i)}{q_b(\mathbf{z}_j)} + q_b(\mathbf{z}_j)\log\frac{q_b(\mathbf{z}_j)}{q_b(\mathbf{z}_i)}\right), \tag{18}$$

where $q_b(\mathbf{z}_i) = \frac{z_{ib}}{\sum_{b=1}^{15B} z_{ib}}$.

Then, based on the posterior class densities $\mathbb{P}(y_i|\mathbf{z}_i)$ and on the label prior $\mathbb{P}(\mathbf{y})$, the MAP results are finally given by

$$\begin{aligned}\hat{\mathbf{y}} &= \underset{\mathbf{y}}{\operatorname{argmax}}\left\{\sum_{i=1}^{N} \log \mathbb{P}(y_i|\mathbf{z}_i) + \beta \sum_{(i,j)\in\mathcal{N}_e} \delta(y_i, y_j)\exp\{-d(\mathbf{z}_i, \mathbf{z}_j)\}\right\},\\ &= \underset{\mathbf{y}}{\operatorname{argmin}}\left\{\sum_{i=1}^{N} - \log \mathbb{P}(y_i|\mathbf{z}_i) - \beta \sum_{(i,j)\in\mathcal{N}_e} \delta(y_i, y_j)\exp\{-d(\mathbf{z}_i, \mathbf{z}_j)\}\right\}\end{aligned} \tag{19}$$

The final label results can be obtained by minimizing (19) applying the efficient α-Expansion graph-cut-based algorithm [23]. In this paper, we denote our classification method using 3D-DWT feature as SVM-3DG. Besides, we denote the classification method using 3D-DWT feature without considering the term $\mathbb{P}(\mathbf{y})$ as SVM-3D. Fig. 2 depicts a flowchart for the proposed classification method and the whole algorithm process is summarized in Algorithm 1.

Algorithm 1. SVM-3DG hyperspectral image classification.

Input: Hyperspectral image data $\mathcal{X} \in R^{H\times W\times B}$, smoothness parameter β, the number of classes K.
Output: Labels $\hat{\mathbf{y}}$.

1: Convert $\mathcal{X}$ into $\widehat{\mathcal{Z}} \in R^{H\times W\times 15B}$ using 3D-DWT, and convert $\widehat{\mathcal{Z}}$ into matrix $\mathbf{Z} = [\mathbf{z}_1; \ldots; \mathbf{z}_{HW}] \in R^{HW\times 15B}$;
2: Randomly select some samples in $\mathbf{Z}$ as training samples and use the remaining as testing ones;
3: Train the probabilistic SVM classifier using training samples;
4: Compute the probabilistic output $\mathbf{P} \in R^{HW\times K}$ (Each row of $\mathbf{P}$ is a normalized vector which exactly interprets the confidential probability of each class) on $\mathbf{Z}$ using the probability SVM;
5: Compute the classification labels $\mathbf{y}$ using α-Expansion ($\mathbf{P}$,β).

5. Experiments

In this section, to validate the effectiveness of our proposed SVM-3D and SVM-3DG methods, we conducted a series of experiments on two benchmark datasets, namely Indian Pines data and Pavia University data. Several state-of-the-art methods were considered for comparison, including standard SVM (SVM), SVM following standard PCA (SVM-PCA), SVM with composite kernels (SVM-CK) [25], graph-cut following SVM (SVM-GC) [22], simultaneous orthogonal matching pursuit (SOMP) [4], subspace multinomial logistic regression with multilevel logistic (MLRsubMLL) [21] and structured sparse logistic regression (SSLR) [7]. All the experiments were implemented in Matlab R2014b on a PC with 3.60 GHz CPU and 16 GB RAM.

All the aforementioned methods are compared numerically using the following three criteria [44]: overall accuracy (OA), average accuracy (AA) and statistically kappa coefficient (κ). Specifically, OA represents the number of correctly classified samples divided by the total number of test samples, AA denotes the average of individual class accuracies, and kappa coefficient (κ) involves both omission and commission errors and accounts for the overall effective performance of the classifier. In order to calculate the three criteria, the classification confusion matrix should be obtained first, which is defined as follows:

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1K} \\ m_{21} & m_{22} & \cdots & m_{2K} \\ & \cdots & \cdots & \\ m_{K1} & m_{K2} & \cdots & m_{KK} \end{pmatrix}, \tag{20}$$

where m_{ij} shows the number of pixels that should belong to class i but are assigned to class j, and K is the number of classification classes. Then, according to the confusion matrix $\mathbf{M}$, OA, AA and kappa coefficient (κ) can be calculated separately as follows. Firstly, the OA

Table 1
Statistics of the Indian Pines data set, including the name, the number of training, test and total samples for each class.

Class		Samples		
No	Name	Train	Test	Total
1	Alfalfa	15	31	46
2	Corn-no till	15	1413	1428
3	Corn-min till	15	815	830
4	Corn	15	222	237
5	Grass-pasture	15	468	483
6	Grass-trees	15	715	730
7	Grass-pasture-mowed	15	13	28
8	Hay-windrowed	15	463	478
9	Oat	15	5	20
10	Soybean-no till	15	957	972
11	Soybean-min till	15	2440	2455
12	Soybean-clean	15	578	593
13	Wheat	15	190	205
14	Woods	15	1250	1265
15	Buildings-Grass-Trees-Drives	15	371	386
16	Stone-Steel-Towers	15	78	93
Total		240	10,009	10,249

can be obtained by

$$OA = \frac{1}{N_{test}} \sum_{i=1}^{K} m_{ii}, \tag{21}$$

where N_{test} is the total number of test samples. Then, the AA is computed as

$$CA_i = \frac{m_{ii}}{N_i},\ i = 1, 2, \ldots, K \tag{22}$$

$$AA = \frac{1}{K} \sum_{i=1}^{K} CA_i, \tag{23}$$

where CA_i is the accuracy of class i and N_i is the total test samples in class i. Finally, the kappa coefficient (κ) is calculated by

$$\kappa = \frac{OA - P_e}{1 - P_e}, \tag{24}$$

where $P_e = \sum_{i=1}^{K} (P_{i.} P_{.i})$ is the expected agreement, $P_{i.} = \frac{R_i}{N_{test}}$, $P_{.i} = \frac{C_i}{N_{test}}$, R_i and C_i represent the sum of ith row and column in confusion matrix, respectively. For all the three criteria, a larger value indicates a better classification performance.

5.1. Classification of Indian pines data

This data set was gathered by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines test site in Northwestern Indiana in June 1992. The original dataset contains 220 spectral reflectance bands in the wavelength range 0.4–2.5 μm, of which 20 bands covering the region of water absorption are removed and thus only 200 bands are reserved in our experiments. This scene has a spectral resolution of 10 nm and a spatial resolution of 20 m by pixel, and the spatial dimension is 145×145. The ground truth contains 16 land cover classes. The number of pixels in each class ranges unbalanced from 20 to 2468, which poses a big challenge for the classification task and is demonstrated in Table 1. A sample band of this image and the related ground truth data are shown in Fig. 3.

Firstly, to evaluate the validity of our proposed SVM-3DG method with small number of labeled samples, we randomly chose 15 samples for each class from the reference data as training samples, and the remaining samples in each class are used to test. This experiment is repeated 20 times. Available training and testing sets are also summarized in Table 1.

Additionally, several parameters need to be tuned in the experiments. For all the SVM-based methods, the RBF parameter γ and the penalty parameter C are tuned through 5-fold cross validation ($\gamma = 2^{-8}, 2^{-7}, \ldots, 2^{8}$, $C = 2^{-8}, 2^{-7}, \ldots, 2^{8}$). Besides, some other parameters of these methods also need to be tuned. Specifically, for SVM-PCA, the number of principal components for PCA is set as 37 in our experiment, which is suggested by [45]. For SVM-CK, the composite weight η is also chosen using 5-fold cross validation ($\eta = 0, 0.1, \ldots, 1$). For SVM-GC, the spatial smoothness parameter β is set as 0.75 advised by [22]. For SOMP, p is set to ∞ in the l_p-norm since it leads to good performance. The dictionary is composed of all of the training samples, and the window size and sparsity level are selected following [4]. For MLRsubMLL, the smoothness parameter μ and the threshold parameter τ are set following [21]. For SSLR, the maximum number of linear sparse classifiers k, the parameter λ in logistic regression and the combinational parameter τ are also set as [7] suggests. For our SVM-3D and SVM-3DG methods, except the aforementioned RBF parameter γ and the penalty parameter C, only the smoothness parameter β needs to be tuned. In our experiments, β is set as 0.75, which was also adopted by SVM-GC.

Classification maps for all the competing methods on Indian dataset are shown in Fig. 4, and the accuracies (i.e. the classification accuracy of each class, OA, AA and kappa coefficient κ) are reported and compared in Table 2. From Fig. 4 and Table 2, one main result can be highlighted: SVM-3DG performs the best in terms of all the three criteria (OA, AA and κ) and attains a large improvement in this scenario, with SVM-GC and SVM-3D trailing slightly behind (76.74% and 72.51% OA, respectively). As shown in Table 2, although only 15 labeled samples for each class (240 samples in total) are selected in the Indian Pines data, the OA of SVM-3DG can reach 81.12%, about 22% higher than that of SVM (59.17%). Moreover, it can be also easily observed that the classification map of SVM-3DG is more closely related to the ground truth map, as depicted in Fig. 4.

We also validate the classification result (OA) by Wilcoxon test [46]. This test is non-parametric statistical hypothesis test, which makes the

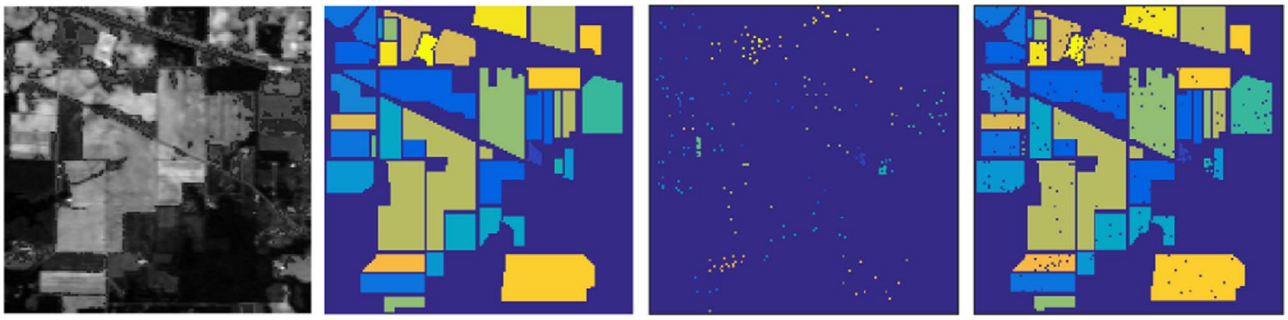


Fig. 3. Indian Pines image and related ground truth categorization information. (a) The original HSI. (b) The ground truth categorization map. (c) The training map. (d) The test map.

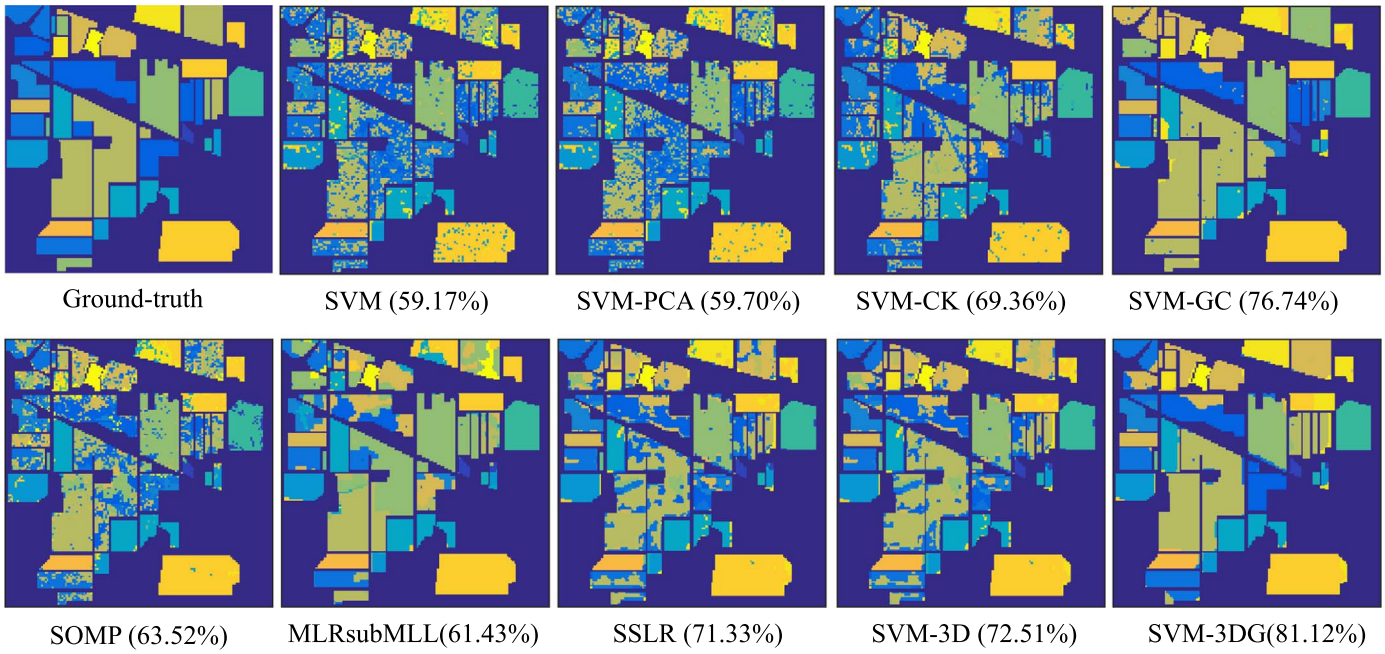


Fig. 4. Classification maps obtained by all competing methods on the Indian Pines dataset (overall accuracies are reported in parentheses).

Table 2
Classification accuracy (%) and standard deviation (in bracket) of all competing methods on the Indian Pines image test set.

Class	SVM	SVM-PCA	SVM-CK	SVM-GC	SOMP	MLRsubMLL	SSLR	SVM-3D	SVM-3DG
1	83.87	83.87	90.32	93.55	83.87	93.55	93.55	93.55	**96.77**
2	40.34	43.74	50.46	55.98	38.71	18.40	47.06	44.80	**58.46**
3	60.49	58.53	64.05	25.89	58.90	70.06	75.46	77.06	**93.37**
4	73.87	77.93	86.94	**97.75**	70.27	52.70	85.64	89.64	96.40
5	77.56	77.56	75.64	83.33	**86.32**	84.19	87.76	83.97	86.11
6	89.09	89.23	83.78	92.03	98.18	**98.88**	89.65	89.65	95.80
7	92.31	92.31	92.31	**100**	**100**	92.31	**100**	**100**	**100**
8	92.87	92.87	96.11	**100**	85.53	**100**	98.24	98.49	**100**
9	**100**	**100**	**100**	0	**100**	**100**	**100**	**100**	**100**
10	69.17	64.99	62.38	74.82	67.08	**77.22**	66.29	65.52	68.86
11	31.56	31.43	59.22	76.19	43.57	37.91	54.67	60.37	**78.57**
12	47.75	50.17	60.90	**97.75**	53.63	69.90	82.01	82.53	96.89
13	95.79	95.79	97.89	97.89	98.42	**100**	94.62	94.74	94.21
14	85.44	88.56	91.68	95.52	92.56	**99.92**	88.60	88.40	77.84
15	51.75	52.02	72.51	83.83	51.75	1.62	86.79	87.06	**95.42**
16	92.31	92.31	97.44	**100**	**100**	**100**	97.44	97.44	98.72
OA	59.17 (0.25)	59.70 (0.21)	69.36 (0.32)	76.74 (0.27)	63.52 (0.35)	61.43 (0.22)	71.33 (0.25)	72.51 (0.24)	**81.12** (0.19)
AA	74.01 (0.23)	74.46 (0.20)	80.10 (0.31)	79.66 (0.25)	76.80 (0.33)	74.79 (0.20)	84.24 (0.26)	84.58 (0.21)	**89.84** (0.18)
κ	54.62 (0.015)	55.15 (0.010)	65.44 (0.020)	73.51 (0.018)	59.20 (0.024)	57.13 (0.011)	67.82 (0.013)	69.11 (0.012)	**78.64** (0.009)
Time (s)	20.26	20.01	3.34	33.14	107.19	5.57	77.97	239.52	348.73

hypothesis decision based on the p-value. Wilcoxon test can determine whether the differences of the classification results between two methods are statistically significant. In Wilcoxon test, if the p-value is smaller than 0.05, the difference between classification accuracies is statistical significant with 95% confidence. The results of Wilcoxon test are shown in Table 3, including comparisons between SVM-3DG and all the other competing methods. The statistical difference ($p-$value < 0.05) in Table 3 indicates that SVM-3DG significantly outperforms other classification methods.

Moreover, from Fig. 4 and Table 2, several observations can also be made. Firstly, SVM and SVM-PCA achieve poorer classification performance compared with other methods, which can also be visually seen from Fig. 4 that the classification maps obtained by both methods contain evidently more salt-and-pepper-like errors. That is because only spectral information is taken into account in both methods, while the other methods consider both spectral and spatial information of HSIs and thus obviously perform better. Secondly, 3-D discrete wavelet transform (3D-DWT) performs better than composite kernels (CK) which can be easily observed by comparing the OA between SVM-3D (72.51%) and SVM-CK (69.36%) in Table 2. Also, it can be visually seen from Fig. 4 that SVM-3D achieves smoother classification map than SVM-CK. Thirdly, SVM-3D and SSLR achieve the third and fourth best

Table 3
p-values obtained through Wilcoxon tests between SVM-3DG and other competing methods on Indian Pines dataset.

Methods	vs. SVM	vs. SVM-PCA	vs. SVM-CK	vs. SVM-GC	vs. SOMP	vs. MLRsubMLL	vs. SSLR	vs. SVM-3D
p-value	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002

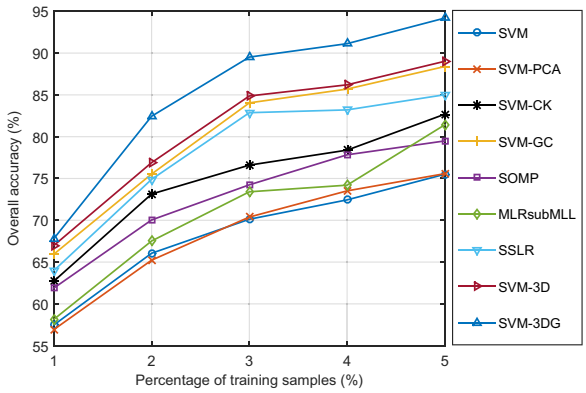


Fig. 5. Overall accuracy (%) of all competing methods with different proportions of training samples on Indian Pine dataset.

performance (72.51% and 71.33%, respectively) due to the use of 3D-DWT feature. Fourthly, by considering the MRF on the neighboring labels, the classification performance can be dramatically improved. As shown in Table 2, the OA of SVM-GC (76.74%) is 17% higher than that of SVM (59.17%) and the OA of SVM-3DG (81.12%) is 9% higher than that of SVM-3D (72.51%). The same phenomenon can also be observed from Fig. 4 that the MRF-based methods, such as SVM-GC, MLRsubMLL and SVM-3DG, provide much smoother classification map. Therefore, the superior performance of our proposed SVM-3DG method can be explained by using the 3D-DWT feature and MRF strategy simultaneously, which fully exploits the spatial information in HSIs. In addition, the execution time of all the methods is also reported in Table 2, from which we can conclude that the methods incorporating 3D-DWT feature and MRF strategy take more time.

In the above experiment, we only consider the case that there are only a small number of training samples. In order to evaluate the effectiveness of our methods with training samples increasing, additional experiments are also conducted, in which 1%, 2%, 3%, 4% and 5% labeled samples of each class from the Indian Pines data are randomly selected as training samples, whereas the rest are taken as testing ones.

The OA of each method is plotted in Fig. 5, from which we can observe that our SVM-3DG method achieves the best performance in all cases. Meanwhile, SVM-3D, SVM-GC and SSLR obtain the second, third and fourth highest OA, respectively. Besides, the OA of SVM-CK is better than SOMP, while SVM and SVM-PCA exhibit the worst performance. All those results are in accordance with the results shown in Table 2.

5.2. Classification of Pavia University data

This dataset was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) over the urban area of the University of Pavia, northern Italy, on July 8, 2002. The original dataset consists of 115 spectral bands ranging from 0.43 to 0.86 μm, of which 12 noisy bands are removed and only 103 bands are retained in our experiments. This scene has a spatial resolution of 1.3 m per pixel, and the spatial dimension is 610×340. There are 9 land cover classes in this scene and the number of each class is displayed in Table 4. Besides, a sample band of this image and the corresponding ground truth map are depicted in Fig. 6.

Table 4
Statistics of the Pavia University data set, including the name, the number of training, test and total samples for each class.

Class		Samples		
No	Name	Train	Test	Total
1	Asphalt	50	6581	6631
2	Meadows	50	18,599	18,649
3	Gravel	50	2049	2099
4	Trees	50	3014	3064
5	Painted metal sheets	50	1295	1345
6	Bare Soil	50	4979	5029
7	Bitumen	50	1280	1330
8	Self-Blocking Bricks	50	3632	3682
9	Shadows	50	897	947
Total		450	42,326	42,776

Firstly, to evaluate the validity of our proposed SVM-3DG method with small number of labeled training samples, we randomly chose 50 samples for each class from the ground truth data as training samples, and the remaining ones for each class are used for testing. This experiment is also repeated 20 times. The related statistics are summarized in Table 4. In this experiment, all parameters involved in the competing methods are tuned in the same way as the previous Indian Pines experiment. Classification maps obtained by all competing methods on Pavia University dataset are displayed in Fig. 6, and the accuracies (i.e. the classification accuracy of each class, OA, AA and kappa coefficient κ) are summarized in Table 5.

From Fig. 4 and Table 2, we can conclude that for this dataset, our SVM-3DG approach achieves the best performance in terms of all the three criteria (OA, AA and kappa coefficient κ). For example, it leads to the highest classification OA (96.06%), even though the number of training sample for each class is only 50, which indicates the superiority of the SVM-3DG in the scenario of small number of labeled training samples. As to SVM, SVM-PCA and SOMP, they achieve comparable classification OA (77.48%, 76.40% and 77.88%, respectively). Also, it is worthy to be emphasized that 3D-DWT performs better than CK, which can be found by comparing the OA of SVM-3D (93.81%) and SVM-CK (83.34%) in Table 5. Meanwhile, it can be visually seen from Fig. 7 that SVM-3D obtains much smoother classification map than SVM-CK. Besides, it should be noted that the SVM-GC achieves 18% higher OA than SVM, and it is also the case that SVM-3DG is 3% higher than SVM-3D, which indicates that the classification accuracy can be obviously improved by adopting the MRF to model the property of label smoothness. In addition, SVM-3D and SSLR achieve good and comparable performance because of the application of 3D-DWT feature. Consequently, the superiority of our SVM-3DG approach can be explained by the use of 3D-DWT and MRF, simultaneously. Moreover, the execution time of each method is also reported in Table 5, from which we can conclude that methods utilizing 3D-DWT feature and MRF strategy cost more time than other ones.

We also validate the classification result (OA) by Wilcoxon test. The results of Wilcoxon test are presented in Table 6, including comparisons between SVM-3DG and all other competing methods. The statistical difference in Table 6 indicates that SVM-3DG achieves a significantly better classification performance than other classification methods ($p-$value < 0.05).

Analogously, to verify the superiority of our methods as training

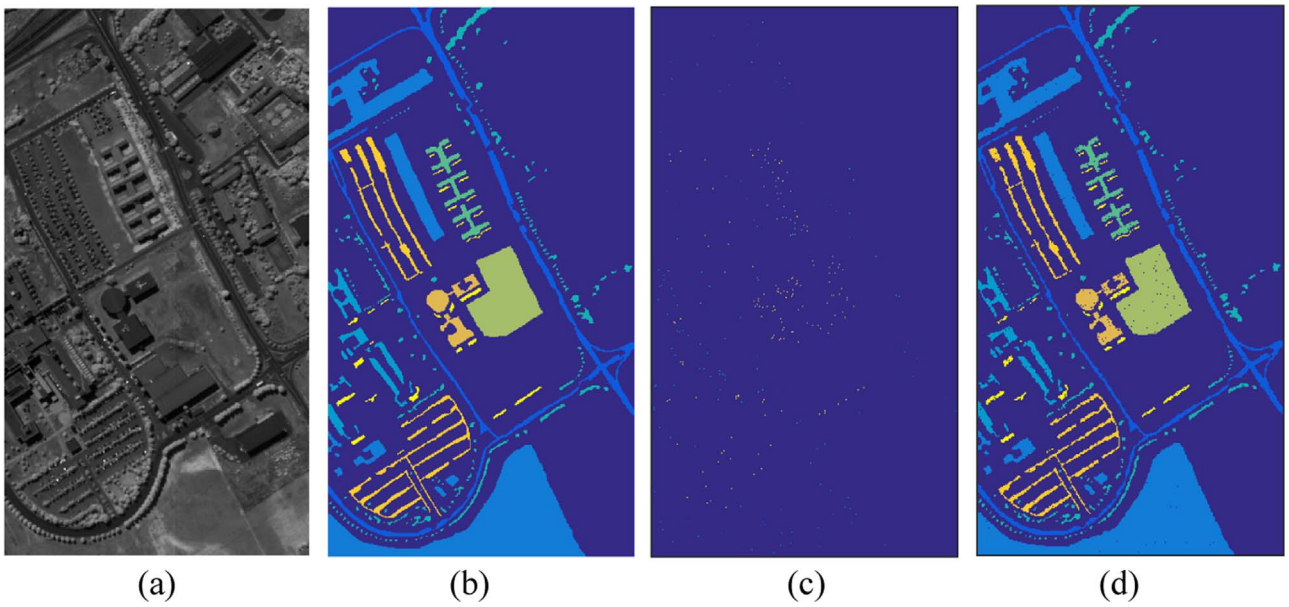


Fig. 6. Pavia University image and related ground truth categorization information. (a) The original HSI. (b) The ground truth categorization map. (c) The training map. (d) The test map.

Table 5
Classification accuracy (%) and standard deviation (in bracket) obtained by all competing methods on the test set of the Pavia University dataset.

Class	SVM	SVM-PCA	SVM-CK	SVM-GC	SOMP	MLRsubMLL	SSLR	SVM-3D	SVM-3DG
1	71.66	71.78	76.74	94.42	48.47	87.62	94.99	94.09	**97.39**
2	79.38	74.85	82.82	95.88	81.98	94.34	90.37	94.92	**97.27**
3	72.67	73.89	77.94	87.12	86.09	77.50	89.12	88.87	**89.41**
4	94.09	94.86	92.40	95.79	93.53	92.57	93.50	96.45	**97.25**
5	99.23	99.46	99.54	99.23	**100**	99.85	99.31	99.61	99.61
6	67.56	71.20	83.19	94.84	74.87	97.31	92.19	95.56	**98.41**
7	82.81	83.59	93.20	94.69	95.55	88.36	93.75	95.63	**98.20**
8	66.55	70.37	80.40	**95.70**	78.28	62.80	81.17	81.64	84.00
9	96.10	95.99	**100**	98.89	99.89	**100**	**100**	99.89	99.89
OA	77.48 (0.75)	76.40 (0.62)	83.34 (0.52)	94.12 (0.55)	77.88 (0.65)	90.10 (0.58)	91.25 (0.56)	93.81 (0.49)	**96.06** (0.47)
AA	81.12 (1.25)	81.78 (1.21)	87.36 (1.32)	95.17 (1.27)	83.75 (1.35)	88.91 (1.22)	92.71 (1.28)	94.07 (1.24)	**95.71** (1.19)
κ	70.87 (0.015)	69.78 (0.011)	78.44 (0.009)	93.68 (0.007)	71.50 (0.007)	86.94 (0.006)	88.55 (0.006)	91.84 (0.006)	**94.78** (0.005)
Time (s)	33.54	29.89	6.90	567.98	1004.56	16.01	415.36	364.94	1672.68

samples increase, we conduct additional experiments, in which 1%, 2%, 3%, 4% and 5% training samples of each class from the Pavia University data are randomly chosen as training samples and the remaining ones are used as testing ones. The OA of each method is shown in Fig. 8, from which we can see that SVM-3DG outperforms the other ones in all variations of training sample size, which is consistent with the result displayed in Table 5.

5.3. Classification in noisy scenario

In this experiment, we evaluate the performance of our SVM-3DG method under the special situation, i.e. noisy scenario. We preserve all of the original 224 spectral bands (i.e. the 200 spectral bands used in the former experiment on Indian Pines data, together with the 24 removed noisy bands) and select the same training (15 for each class) and testing samples as the former Indian Pines experiment did. The OA of all the methods on clean Indian Pines (with 24 noisy bands removed) and noisy Indian Pines are compared in Fig. 9, respectively. From Fig. 9, it can be seen that although the classification performance of SVM-3DG on noisy Indian Pines is worse than that of the experiment on clean Indian Pines, SVM-3DG still achieves higher OA than other competing methods in this noisy scenario, which verifies the better robustness of the proposed SVM-3DG method beyond others in the noisy scenario.

5.4. Impact of parameters

In our SVM-3DG method, except the RBF parameter γ and the penalty parameter C, which are tuned through 5-fold cross validation, the smoothness parameter β is also important to the final performance. In this section, we investigate the performances of SVM-3DG in term of OA with different smoothness parameter $\beta = [0.25, 0.5, 0.75, 1, 2, 3, 4, 5]$. The experimental results are demonstrated in Fig. 10. It can be observed that the OA is not significantly different over the given βs and our method is robust to the selection of β. Therefore, throughout all our experiments, we just easily set β as

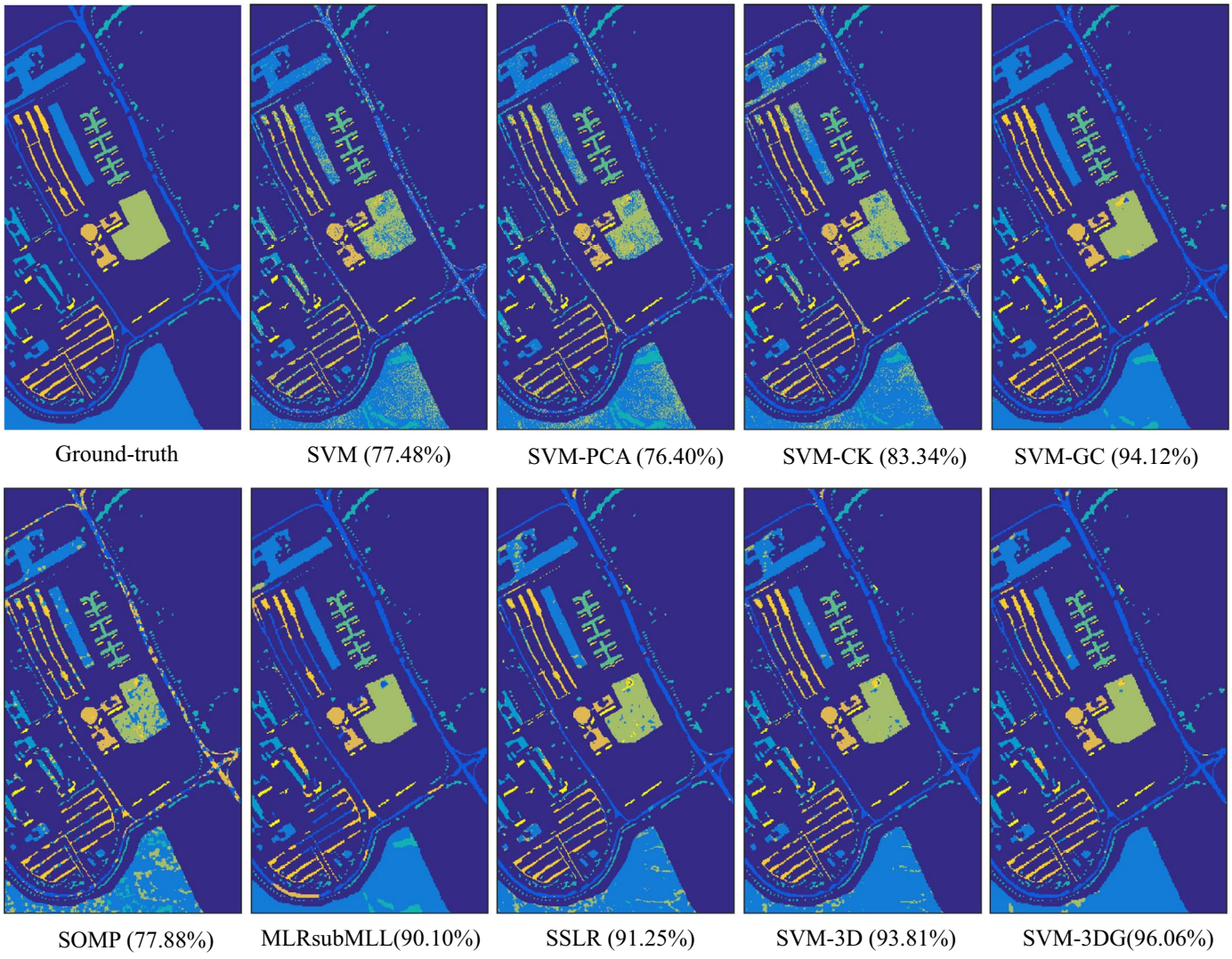


Fig. 7. Classification maps obtained by all competing methods on the Pavia University dataset (overall accuracies are reported in parentheses).

Table 6
p-values obtained through Wilcoxon tests between SVM-3DG and other competing on Pavia University dataset.

Methods	vs. SVM	vs. SVM-PCA	vs. SVM-CK	vs. SVM-GC	vs. SOMP	vs. MLRsubMLL	vs. SSLR	vs. SVM-3D
p-value	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002

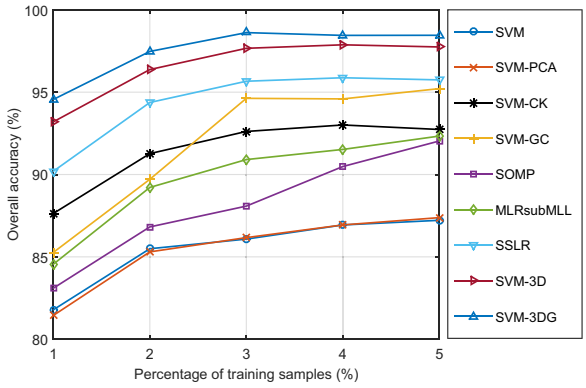


Fig. 8. Overall accuracy (%) obtained by all competing methods with different proportions of training samples on Pavia University dataset.

0.75 and our method can perform consistently well in all cases, as substantiated in the above sections.

6. Conclusions

In this paper, a novel technique is proposed for HSI classification. The key idea is to simultaneously utilize spatial-filtering method and spatial smoothness prior of labels. In this way, the spatial correlation under HSIs can be fully discovered. The 3-dimensional discrete wavelet transform (3D-DWT) is utilized to extract the spectral-spatial features and the probabilistic SVM is used to character the spectral information of the new version of HSI cube reformulated by 3D-DWT. The local correlation of neighboring pixels is further encoded by Markov random field (MRF). By computing the MAP classification with an optimized α-Expansion graph-cut-based algorithm, the proposed classification method is efficient. Experimental results on two real benchmark hyperspectral data sets show that our method outperforms the state-

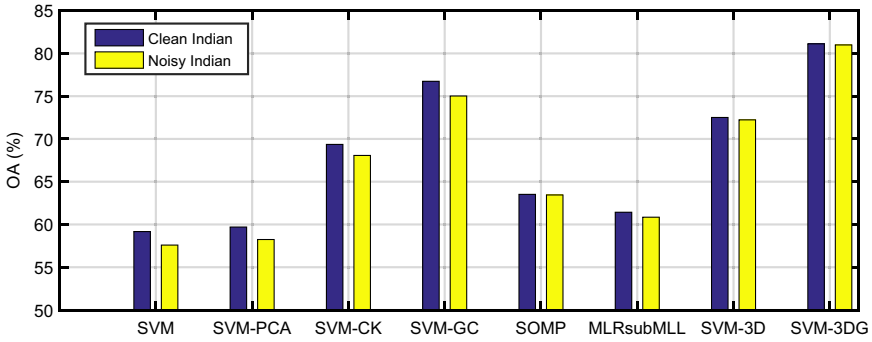


Fig. 9. Overall accuracy (%) of all competing methods in clean and noisy Indian Pines datasets.

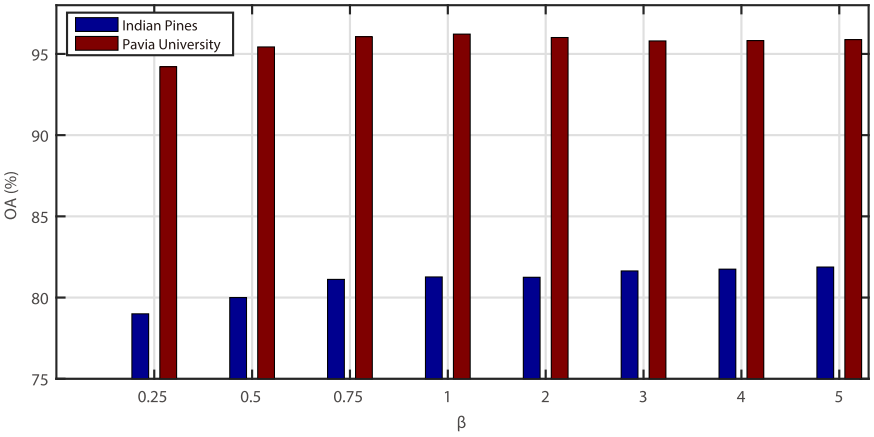


Fig. 10. Overall accuracy (%) obtained under different values of β in the proposed methods for two utilized datasets.

of-the-art methods investigated in this paper. In our future research, we will analyze in depth the pros and cons of our method for classifying different types of classes and attempt to explore more global correlations for feature extraction.

Acknowledgements

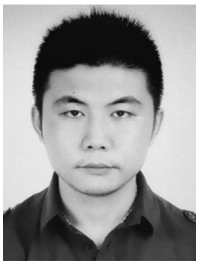
This research was supported by the National Grand Fundamental Research 973 Program of China under Grant no. 2013CB329404 and the China NSFC project under Contract 11131006 and 61373114.

References

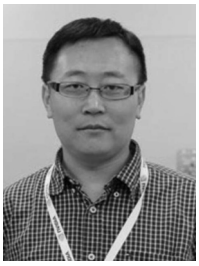
[1] D.A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing 29, John Wiley & Sons, Hoboken, New jersey, 2005.
[2] F. Melgani, L. Bruzzone, Classification of hyperspectral remote sensing images with support vector machines, IEEE Transactions on Geoscience and Remote Sensing 42 (8) (2004) 1778–1790.
[3] J. Ham, Y. Chen, M.M. Crawford, J. Ghosh, Investigation of the random forest framework for classification of hyperspectral data, IEEE Transactions on Geoscience and Remote Sensing 43 (3) (2005) 492–501.
[4] Y. Chen, N.M. Nasrabadi, T.D. Tran, Hyperspectral image classification using dictionary-based sparse representation, IEEE Transactions on Geoscience and Remote Sensing 49 (10) (2011) 3973–3985.
[5] R. Ji, Y. Gao, R. Hong, Q. Liu, D. Tao, X. Li, Spectral-spatial constraint hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 52 (3) (2014) 1811–1824.
[6] Z. He, L. Liu, R. Deng, Y. Shen, Low-rank group inspired dictionary learning for hyperspectral image classification, Signal Processing 120 (2016) 209–221.
[7] Y. Qian, M. Ye, J. Zhou, Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features, IEEE Transactions on Geoscience and Remote Sensing 51 (4) (2013) 2276–2291.
[8] R.D. Phillips, C.E. Blinn, L.T. Watson, R.H. Wynne, An adaptive noise-filtering algorithm for aviris data with implications for classification accuracy, IEEE Transactions on Geoscience and Remote Sensing 47 (9) (2009) 3168–3179.
[9] Y. Chen, N.M. Nasrabadi, T.D. Tran, Hyperspectral image classification via kernel sparse representation, IEEE Transactions on Geoscience and Remote Sensing 51 (1) (2013) 217–231.
[10] A. Soltani-Farani, H.R. Rabiee, S.A. Hosseini, Spatial-aware dictionary learning for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 53 (1) (2015) 527–541.
[11] X. Sun, Q. Qu, N.M. Nasrabadi, T.D. Tran, Structured priors for sparse-representation-based hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 11 (7) (2014) 1235–1239.
[12] Y. Xu, Z. Wu, Z. Wei, Spectral-spatial classification of hyperspectral image based on low-rank decomposition, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8 (6) (2015) 2370–2380.
[13] X. Cao, Y. Chen, Q. Zhao, D. Meng, Y. Wang, D. Wang, Z. Xu, Low-rank matrix factorization under general mixture noise distributions, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1493–1501.
[14] Q. Zhao, D. Meng, Z. Xu, W. Zuo, L. Zhang, Robust principal component analysis with complex noise, in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 55–63.
[15] D. Meng, F. Torre, Robust matrix factorization with unknown noise, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1337–1344.
[16] D. Meng, Z. Xu, L. Zhang, J. Zhao, A cyclic weighted median method for l1 low-rank matrix factorization with missing entries, in: Proceedings of the Association for the Advance of Artificial Intelligence, Vol. 4, 2013, p. 6.
[17] Q. Xie, Q. Zhao, D. Meng, Z. Xu, S. Gu, W. Zuo, L. Zhang, Multispectral images denoising by intrinsic tensor sparsity regularization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1692–1700.
[18] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, B. Zhang, Decomposable nonlocal tensor dictionary learning for multispectral image denoising, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2949–2956.
[19] S. Jia, X. Zhang, Q. Li, Spectral-spatial hyperspectral image classification using regularized low-rank representation and sparse representation-based graph cuts, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8 (6) (2015) 2473–2484.
[20] Y. Tarabalka, J.A. Benediktsson, J. Chanussot, Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques, IEEE Transactions on Geoscience and Remote Sensing 47 (8) (2009) 2973–2987.
[21] J. Li, J.M. Bioucas-Dias, A. Plaza, Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields, IEEE Transactions on Geoscience and Remote Sensing 50 (3) (2012) 809–823.
[22] Y. Tarabalka, A. Rana, Graph-cut-based model for spectral-spatial classification of hyperspectral images, in: Proceedings of the 2014 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2014, pp. 3418–3421.
[23] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2001) 1222–1239.
[24] M. Weeks, M.A. Bayoumi, Three-dimensional discrete wavelet transform architectures, IEEE Transactions on Signal Processing 50 (8) (2002) 2050–2063.

[25] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Calpe-Maravilla, Composite kernels for hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 3 (1) (2006) 93–97.
[26] Y. Gu, Q. Wang, H. Wang, D. You, Y. Zhang, Multiple kernel learning via low-rank nonnegative matrix factorization for classification of hyperspectral imagery, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8 (6) (2015) 2739–2751.
[27] E. Zhang, X. Zhang, L. Jiao, H. Liu, S. Wang, B. Hou, Weighted multifeature hyperspectral image classification via kernel joint sparse representation, Neurocomputing 178 (2016) 71–86.
[28] H. Zhang, J. Li, Y. Huang, L. Zhang, A nonlocal weighted joint sparse representation classification method for hyperspectral imagery, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (6) (2014) 2056–2065.
[29] D. Böhning, Multinomial logistic regression algorithm, Annals of the Institute of Statistical Mathematics 44 (1) (1992) 197–200.
[30] S. Yu, S. Jia, C. Xu, Convolutional neural networks for hyperspectral image classification, Neurocomputing.
[31] R. Moreno, F. Corona, A. Lendasse, M. Graña, L.S. Galvão, Extreme learning machines for soybean classification in remote sensing hyperspectral images, Neurocomputing 128 (2014) 207–216.
[32] C. Qi, Z. Zhou, Y. Sun, H. Song, L. Hu, Q. Wang, Feature selection and multiple kernel boosting framework based on pso with mutation mechanism for hyperspectral classification, Neurocomputing.
[33] M. Volpi, G. Matasci, M. Kanevski, D. Tuia, Semi-supervised multiview embedding for hyperspectral data classification, Neurocomputing 145 (2014) 427–437.
[34] P. Zhong, R. Wang, Learning conditional random fields for classification of hyperspectral images, IEEE Transactions on Image Processing 19 (7) (2010) 1890–1907.
[35] Y. Tarabalka, J. Chanussot, J.A. Benediktsson, Segmentation and classification of hyperspectral images using watershed transformation, Pattern Recognition 43 (7) (2010) 2367–2379.
[36] L. Chun-Lin, A tutorial of the wavelet transform, NTUEE, Taiwan.
[37] S. Geman, D. Geman, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell., 6,1984, pp. 721–741.
[38] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 27.
[39] T.-F. Wu, C.-J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, The Journal of Machine Learning Research 5 (2004) 975–1005.
[40] Y. Tarabalka, M. Fauvel, J. Chanussot, J.A. Benediktsson, Svm-and mrf-based method for accurate classification of hyperspectral images, IEEE Geoscience and Remote Sensing Letters 7 (4) (2010) 736–740.
[41] M. Khodadadzadeh, J. Li, A. Plaza, H. Ghassemian, J. M. Bioucas-Dias, Spectral-spatial classification for hyperspectral data using svm and subspace mlr, in: Proceedings of the 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2013, pp. 2180–2183.
[42] W. Liao, J. Tang, B. Rosenhahn, M. Y. Yang, Integration of gaussian process and mrf for hyperspectral image classification, in: Proceedings of the 2015 Joint Urban Remote Sensing Event (JURSE), IEEE, 2015, pp. 1–4.
[43] G. Moser, S.B. Serpico, J.A. Benediktsson, Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images, Proceedings of the IEEE 101 (3) (2013) 631–651.
[44] L. Wang, C. Zhao, Hyperspectral Image Processing, Springer, Berlin, German, 2015.
[45] H. Li, G. Xiao, T. Xia, Y.Y. Tang, L. Li, Hyperspectral image classification using functional data analysis, IEEE Transactions on Cybernetics 44 (9) (2014) 1544–1555.
[46] J.D. Gibbons, S. Chakraborti, Nonparametric Statistical Inference, Springer, Berlin, German, 2011.

Xiangyong Cao received the B.Sc. degree in 2012 from Xi'an Jiaotong University, Xi'an, China. He is currently pursuing the Ph.D. degree at Xi'an Jiaotong University. His current researchinterests include low-rank modeling, statistical modeling and hyperspectral image analysis.



Lin Xu is currently pursuing the Ph.D. degree with the Institute for Information and System Sciences, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. His current research interests include neural networks, learning algorithms, and applications in computer vision.



Deyu Meng received the B.Sc., M.Sc., and Ph.D. degrees in 2001, 2004, and 2008, respectively, from Xi'an Jiaotong University, Xi'an, China.

He is currently an Associate Professor with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University. From 2012 to 2014, he took his two-year sabbatical leave in Carnegie Mellon University. His current research interests include self-paced learning, noise modeling, and tensor sparsity.

Qian Zhao received the B.Sc. and Ph.D degrees from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2015, respectively.

He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2013 to 2014. He is currently a Lecturer with School of Mathematics and Statistics, Xi'an Jiaotong University. His current research interests include low-matrix/tensor analysis, Bayesian modeling and self-paced learning.

Zongben Xu received the Ph.D. degree in mathematics-from Xi'an Jiaotong University, China, in 1987. He serves as the Chief Scientist of National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences with the Xi'an Jiaotong University. His current research interests include intelligent information processing and applied mathematics.

Dr. Xu was a recipient of the National Natural Science Award of China in2007 and was a winner of the CSIAM Su Buchin Applied Mathematics Prizein 2008. He delivered a talk at the International Congress of Mathematicians 2010. He was elected as a member of Chinese Academyof Science in 2011.