Scene Classification Based on the Fully Sparse Semantic Topic Model

Qiqi Zhu, Yanfei Zhong, *Senior Member, IEEE*, Liangpei Zhang, *Senior Member, IEEE*, and Deren Li, *Senior Member, IEEE*

***Abstract*— In high spatial resolution (HSR) imagery scene classification, it is a challenging task to recognize the high-level semantics from a large volume of complex HSR images. The probabilistic topic model (PTM), which focuses on modeling topics, has been proposed to bridge the so-called semantic gap. Conventional PTMs usually model the images with a dense semantic representation and, in general, one topic space is generated for all the different features. However, this approach fails to consider the sparsity of the semantic representation, the classification quality, as well as the time consumption. In this paper, to solve the above problems, a fully sparse semantic topic model (FSSTM) framework is proposed for HSR imagery scene classification. FSSTM, with an elaborately designed modeling procedure, is able to represent the image with sparse but representative semantics. Based on this framework, the topic weights of multiple features are exploited by solving a concave maximization problem, which improves the fusion of the discriminative semantic information at the topic level. Meanwhile, the sparsity and representativeness of the topics generated by FSSTM guarantee that the image is adaptive to the change of a topic number. FSSTM can consistently achieve a good performance with a limited number of training samples, and is robust for HSR image scene classification. The experimental results obtained with three different types of HSR image data sets confirm that the proposed algorithm is effective in improving the performance of scene classification, and is highly efficient in discovering the semantics of HSR images when compared with the state-of-the-art PTM methods.**

***Index Terms*— Fusion, high spatial resolution (HSR) imagery, limited training samples, probabilistic topic model (PTM), scene classification, sparsity.**

I. Introduction

A LARGE amount of high spatial resolution (HSR) images with abundant spectral and spatial information are now available. With a spatial resolution of up to half a meter, the HSR images can enable more accurate surface observation. Object-based and contextual-based methods have been widely applied to HSR images, and can achieve precise object recognition [1], [2]. However, an HSR scene is often composed of diverse objects, such as buildings, trees, and roads. Even when land-use objects or regions can be recognized correctly, manual interpretation of the high-level semantics of the scenes is often essential for many applications of HSR image analysis, such as ecological analysis in public health studies, land-cover classification, and urban mapping and monitoring [3]. The same scenes may contain different land-cover objects, and the same type of objects may differ in the spectral, textural, and structural characteristics. Different spatial arrangements of the same type of objects may also lead to different scenes. Based on the low-level features, the traditional methods are unable to capture the complex semantic concepts of HSR images. This leads to a divergence between the low-level data and the high-level semantic information, namely, the so-called semantic gap [4]. According to the geographical properties, scene classification can be used to automatically label an HSR image to obtain regions with different semantic informations.

Scene classification methods based on semantic object recognition and local visual words are used to capture the high-level semantics of HSR images [8]–[10]. On the other hand, deep learning approaches have consistently presented their superiority in remote sensing tasks [5]–[7]. However, these methods usually require a large number of training samples, high time consumption, and high-level hardware. Hu *et al.* [46] transferred features from a pretrained deep convolutional neural network (CNN) for HSR image scene classification. The gradient-boosting CNN framework, which effectively combines different deep neural networks, is also effective for HSR scene classification [47]. Among the diverse scene classification methods, the bag-of-visual-words (BOVW) model [12]–[14], as an intermediate feature representation method, has been successfully applied to capture the high-level information of HSR scenes, without the recognition of objects. The probabilistic topic models (PTMs), such as probabilistic latent semantic analysis (PLSA) [15] and latent Dirichlet allocation (LDA) [16], [17], mine the latent topics from the images to represent the scenes as a random mixture of visual words. PTMs have been successfully employed to solve the challenges of HSR image scene classification [18]–[21].

In the previous PTM-based scene classification approach, LDA is utilized as both the feature extractor and scene classifier [19], where an image is assigned to the class that maximizes the likelihood inferred by LDA. However, in this way, multiple models are built for the different scene classes, and the correlations between the latent topic spaces of different scene classes are not considered. For the LDA model, i.e.,

Manuscript received November 24, 2016; revised January 7, 2017 and April 15, 2017; accepted May 4, 2017. Date of publication June 26, 2017; date of current version September 25, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0504202, in part by the National Natural Science Foundation of China under Grant 41622107 and Grant 41371344, and in part by the Natural Science Foundation of Hubei Province under Grant 2016CFA029. *(Corresponding author: Yanfei Zhong.)*

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: wuxi5477@126.com; zhongyanfei@whu.edu.cn; zlp62@whu.edu.cn; drli@whu.edu.cn).

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.2017.2709802

SAL-LDA [22], the distribution of the topic variables in LDA is drawn from a Dirichlet distribution with the parameter, and the topic variable is greater than 0, no matter how α varies [16]. To acquire more semantic information from a large number of images, the number of topics may have to be increased. However, this dense semantic representation usually contains a lot of unrepresentative information for the scene, which requires more storage space and is time-consuming with the complex topic modeling. To solve this problem, some researchers have imposed sparsity constraints on the topics to change the objective function of the model [26], [27]. However, these models usually require model selection with many regularization term-based auxiliary parameters, which may be problematic with large-scale data sets.

In addition, the performance of the popular scene classification methods is usually heavily dependent on the number and the quality of the training samples. Fivefold cross validation is often performed to evaluate an experimental data set, to guarantee enough training samples for the HSR image scene classification [8], [9]. However, the annotation of the training samples may be impossible due to many real-world problems, and usually costs a lot of time [31]. It is, therefore, expected that only a few training samples are used to train the classification model, which is important in practice for limited training samples.

The fully sparse topic model (FSTM) [28], which does not employ the Dirichlet prior, was proposed for the modeling of large collections of documents with a sparse latent representation. To date, in addition to the use in natural language analysis, FSTM has been applied to content-based video retrieval [29] and abnormality detection in traffic videos. FSTM is able to discover the latent motion patterns in the video to detect abnormal events [30]. However, when directly using FSTM for scene classification, it is unable to generate discriminative semantics for the HSR images. Multifeature-based scene classification methods have been proposed and are effective in discriminating the semantic information of different scenes [23]–[25]. In general, one dictionary and one semantic space are acquired for the multiple features, because the different features are usually fused before k-means clustering. As the k-means algorithm is not efficient in the high-dimensional feature space, this approach is unable to circumvent the inadequate clustering capacity of the hard-assignment-based k-means algorithm [14]. This also leads to mutual interference between the multiple features.

In this paper, we simultaneously tackle these problems by proposing the fully sparse semantic topic model (FSSTM). First, the heterogeneous features are extracted from the scenes, and are quantized separately to circumvent the hard-assignment effect of k-means clustering. The FSSTM is then employed to mine the sparse latent topics from the low-level feature representation. The weights of the different types of features are optimized by solving a concave maximization problem in FSSTM, which leads to improved fusion of the multiple types of sparse topics. Derived from the sparse prior, the sparse and heterogeneous topics are fused at the topic level, and a generative/discriminative hybrid strategy is used to classify the scenes. The proposed generative model, FSSTM, is used to extract the topic features, and the discriminative classifier, support vector machine (SVM) with a histogram intersection kernel (HIK), is effective in increasing the discrimination of different scenes. FSSTM with sparse semantic distribution changes little when the topic number set is manually increased. This is consistent with the fact that an image can usually be explained by only a few representative topics, which contribute a lot to the semantic understanding of the image. This implies that FSSTM is able to robustly discover compact and effective semantics from an image.

The main contributions of this paper are presented as follows.

1) We propose the FSSTM framework to discover intrinsic low-level information from the image, and represent complex HSR images with sparse topics. To capture the important correlation between the multiple topic spaces of various scene classes, a generative/discriminative hybrid strategy is utilized. In this way, we can both speed up the traditional PTMs, i.e., LDA and FSTM, and extract powerful high-level semantics for the subsequent scene classifier.
2) An effective procedure is elaborately designed for scene classification. The different types of features are converted into heterogeneous 1-D histograms based on multiple dictionaries, and multiple latent topic spaces are proposed for the holistic scene representation of HSR images. The inference task of optimization is reformulated as a concave maximization problem, which leads to improved fusion of the multiple types of topics, and thus, a sparse but representative semantic description is obtained for HSR scenes.
3) We explore a robust PTM, which is adaptive to the change of the number of training samples. FSSTM presents a better performance than the existing relevant methods, even with a limited number of training samples, and is more in line with practical applications. The time consumption is greatly reduced compared with the conventional FSTM and the classical PTM. We performed extensive experiments to verify the accuracy and efficiency of the proposed method with diverse data sets, including an aerial image and two satellite images.

The rest of this paper is organized as follows. Section II introduces the background to the PTM. Section III details the procedure of the proposed FSSTM for HSR image scene classification. A description of the experimental data sets and an analysis of the experimental results are presented in Section IV. A sensitivity analysis is given in Section V. Finally, the conclusions are provided in Section VI.

II. Background

The BOVW model has been a popular approach for content-based image classification due to its simplicity and good performance. The methods using the BOVW model are strongly reliant on the extraction of the low-level features, which are the aggregation of the local image information. However, the investigation of strong discriminative power and invariant prosperities to geometrical changes in the complex HSR images

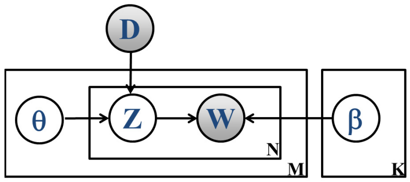

Fig. 1. Probabilistic graphical model of PLSA.

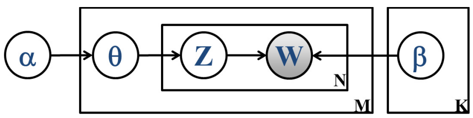

Fig. 2. Probabilistic graphical model of LDA.

needs much prior knowledge and expertise in related fields. This is difficult and time-consuming, especially in the case of massive images [32].

A. Classical Probabilistic Topic Models

Based on the BOVW model, the PLSA model was proposed by Hofmann [15] in 2001. As shown in Fig. 1, the nodes D, Z, and W represent the image, the topic, and the visual word, respectively. The index of the N local patches is denoted as j, the index of M images is denoted as i, and the index of the topics is denoted as k. By choosing a topic z_k with probability $\theta = p(z_k|d_i)$ from K topics, and a word w_j with probability $\beta = p(w_j|z_k)$, the probability $p(w_j|d_i)$ between visual words w_j and images d_i can be decomposed as (1)

$$p(w_j|d_i) = \sum_{k=1}^{K} p(w_j|z_k)p(z_k|d_i). \tag{1}$$

However, PLSA lacks a probability function to describe the images, which means that the number of model parameters grows linearly with the size of the images [44].

In 2003, Blei *et al.* [16] proposed LDA, which introduces the Dirichlet distribution to the topic mixture θ based on the PLSA model, as shown in Fig. 2. The K-dimensional random variable $\theta = \{\theta_1, \ldots, \theta_i, \ldots, \theta_M\}$, where $\theta_i = \{\theta_{i1}, \ldots, \theta_{ik}, \ldots, \theta_{iK}\}$, follows a Dirichlet distribution with parameter $\alpha = \{\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K\}$. LDA provides a probability function for the discrete latent topics in PLSA, and is, therefore, a complete PTM.

LDA has been widely applied to spatial analysis [20], semantic object clustering [33], and HSR image scene classification [19], [22]. However, the Dirichlet variable is greater than 0 when varies. The latent semantics mined from images by LDA are dense when a large amount of images are modeled, which results in a lot of useless information and requires a lot of storage space. The topic modeling is, therefore, complex and takes a lot of time.

B. Fully Sparse Topic Model

In 2012, FSTM was proposed by Than and Ho [28] and was applied to supervised dimension reduction. FSTM is a simplified variant of PLSA and LDA. It removes the endowment of the Dirichlet distribution of LDA, and deliberately allows only a few topics to contribute to an image. It is also a variant of PLSA when removing the observed variable associated with each document. In more detail, FSTM assumes that an image is represented as K topics, denoted by $\beta_1, \ldots \beta_K$. Given a data set consisting of M images $\mathbf{d}$, each image can be described by a set of N visual words w_j.

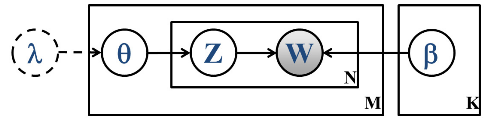

Fig. 3. Probabilistic graphical model of FSTM.

The generative process with FSTM follows two steps for each image.

1) Randomly choose a K-dimensional topic proportion θ.
2) For the jth word in $\mathbf{d}$, the following hold.
   a) Choose a latent topic z_k with probability $P(z_k|\mathbf{d}) = \theta_k$.
   b) Generate a word w_j with probability $P(w_j|z_k) = \beta_{kj}$.

Despite being a simplified variant, FSTM has many interesting properties. Even though there is no explicit prior over the topic proportions, it was shown in [28] that an implicit prior does in fact exist. The implicit prior conforms to the density function

$$p(\theta|\lambda) \propto \exp(-\lambda \cdot \|\theta\|_0) \tag{2}$$

where $\|\theta\|_0$ is the number of nonzero entries of θ. The latent topic proportion θ in FSTM follows an implicit constraint $\|\theta\|_0 \leq L + 1$, where L is the iteration times. This property is a consequence of the sparse inference in FSTM. It is termed implicit modeling, and it allows FSTM to be able to converge at a linear rate to the optimal solutions. Hence, we chose FSTM with sparse solutions to model the HSR imagery in this paper. FSTM utilizes a graphical model to represent the relationship between the image, topic, and visual word, as shown in Fig. 3. However, when using FSTM to discover sparse information, it is unable to capture the complex spatial arrangements of HSR images when it is directly applied to scene classification.

III. Scene Classification Based on the Fully Sparse Semantic Topic Model

To effectively utilize sparse but representative semantic information, the FSSTM framework is proposed for HSR image scene classification. Four tasks have to be addressed for the scene classification (Fig. 4): 1) extracting heterogeneous low-level features from each image patch to describe the complex image; 2) mining sparse but adequate latent semantics using the FSSTM model; and 3) latent semantic fusion and classification with the SVM classifier. The overall flowchart of scene classification based on the FSSTM model is shown in Fig. 4.

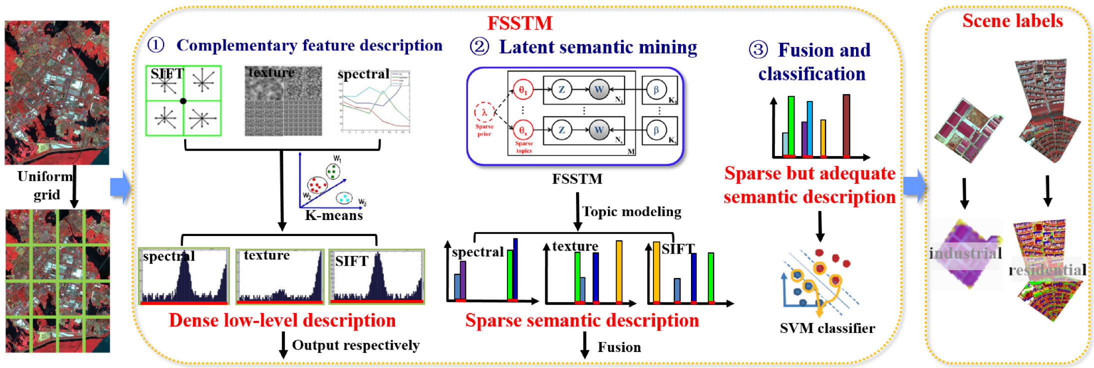

Fig. 4. Flowchart of scene classification based on the FSSTM model for HSR imagery.

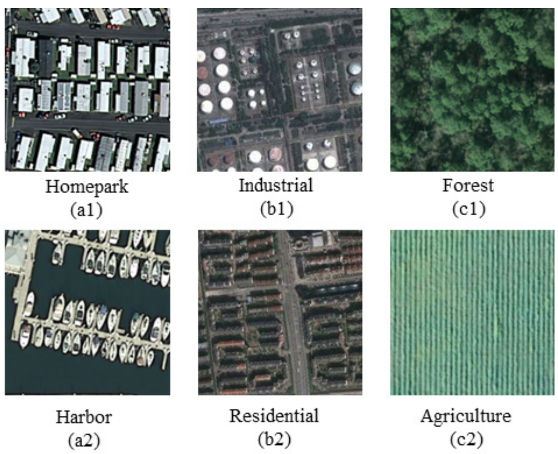

Fig. 5. HSR images of the parking lot, harbor, storage tanks, dense residential, forest, and agriculture scene classes. (a1) and (a2) Importance of the spectral characteristics for HSR images. (b1) and (b2) Importance of the structural characteristics for HSR images. (c1) and (c2) Importance of the textural characteristics for HSR images.

A. Heterogeneous Feature Description

In Fig. 5(a1) and (a2), the home park and harbor scene classes are similar in the structural and textural characteristics, and the spectral feature is most effective in distinguishing them. From Fig. 5(b1) and (b2), it can be seen that the industrial scene and residential scene differ in their structural characteristics. In addition, it is clear that the forest scene and agriculture scene differ in their textural characteristics. To capture the distinctive characteristics of the complex scenes, three complementary features are designed for the HSR imagery scene classification task. Before the feature descriptor extraction, the images are split into image patches using the uniform grid sampling method.

1) The spectral feature reflects the attributes that constitute the ground components and structures. The first-order statistics of the mean value and the second-order statistics of the standard deviation value of the image patches are calculated in each spectral channel as the spectral feature. According to

$$\mathrm{mean}_j = \frac{\sum_{i=1}^{n} v_i}{n} \tag{3}$$

$$\mathrm{std}_j = \sqrt{\frac{\sum_{i=1}^{n} (v_{ij} - \mathrm{mean}_j)^2}{n}} \tag{4}$$

n is the total number of image pixels in the sampled patch, and v_{ij} denotes the jth band value of the ith pixel in a patch. In this way, the mean (mean_j) and standard deviation (std_j) of the spectral vector of the patch are then acquired.

2) The texture (TEX) feature contains information about the spatial distribution of the tonal variations within a band [34], which can give consideration to both the macroscopic properties and fine structure. Wavelet transforms enable the decomposition of the image into different frequency subbands, which is similar to the way the human visual system operates [35]. This makes it especially suitable for image classification. Multilevel 2-D wavelet decomposition is utilized to capture the TEX feature from the HSR images. The level of the wavelet decomposition for the images is optimally set to 3.

3) The scale-invariant feature transform (SIFT) feature [36] has been widely applied in image analysis, since it can overcome the addition of noise, affine transformation, and changes in illumination, as well as compensating for the deficiency of the spectral feature for HSR imagery. Each image patch is split into neighborhood regions, and each direction for each gradient orientation histogram is counted in each region. Hence, the gray dense SIFT descriptor with 128 dimensions is extracted as the structural feature. This was inspired by previous work, in which dense features performed better for scene classification [37]. In addition, Lowe [36] suggested that using a $4 \times 4 \times 8 = 128$-dimension vector to describe the key-point descriptor is optimal.

B. Latent Semantic Mining Based on FSSTM

The previous studies have shown that a uniform grid sampling method can be more effective than other sampling methods, such as random sampling [37]. As a PTM, FSSTM utilizes a visual analog of a word, acquired by vector quantizing the region descriptors [18]. In this way, the image patches acquired by uniformly sampling the HSR images are digitized by S types of features, and all the types of feature descriptors, $D_1, \ldots, D_S$, are obtained. However, with the influence of illumination, rotation, and scale variation, the same visual word in different images may be endowed with various feature

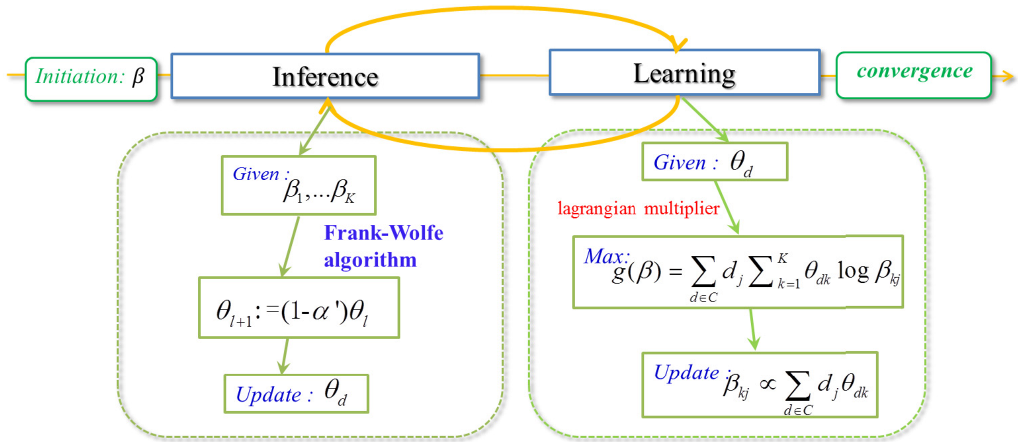

Fig. 6. Algorithm inference and learning task of FSSTM for one type of feature.

values. The k-means clustering is applied to quantize the feature descriptors to generate a 1-D frequency histogram, where image patches with similar feature values correspond to the same visual word. By statistical analysis of the frequency of each visual word, we can obtain the corresponding visual vocabulary.

The conventional methods usually directly concatenate the S types of feature descriptors to make up a long feature $F_1 = \{D_1, \ldots, D_S\}$, which is named concatenate (CAT)-FSTM. The long vector is then quantized by k-mean clustering to generate a 1-D histogram for all the features. However, as the features interfere with each other when clustering, the 1-D histogram is unable to fully describe the HSR imagery. Accordingly, FSSTM is designed to avoid the shortcoming of the conventional methods.

As a statistical method, FSSTM analyzes the visual words of the original images to discover the topics that run through them, how these topics are connected to each other, and how they change over time [17]. By introducing the topics characterized by a distribution over words, FSSTM models the images as random mixtures over the latent variable space. Labeling or annotation of the images is not required in FSSTM. The scene classification methods using FSSTM are able to greatly reduce the dimension of the feature vectors for the representation of the HSR images. In FSSTM, the S types of features are quantized separately by the k-mean clustering algorithm to acquire S distinct 1-D histograms, $H_1, \ldots, H_S$. By introducing probability theory, each element of the 1-D histogram for FSSTM is transformed into the word occurrence probability. To mine the most discriminative semantic features, which is also the core idea of the PTM, the S histograms are separately mined by FSSTM to generate S distinct latent topic spaces. This is different from the conventional strategies which fuse the S histograms before topic modeling, obtaining only one latent topic space, which is inadequate.

Hence, given an image data set $\mathbf{C}$, the dth image $\mathbf{M}$ can be described by a set of words w_j. Hence, an image can be represented as $M = \{w_1, \ldots, w_j, \ldots, w_N\}$, where $w_j \in \{1, 2, \ldots, V\}$, and V is the number of visual words from the visual dictionary. Specifically, FSSTM chooses a k-dimensional latent variable θ. Then, for each of the $H_1, \ldots, H_S$ values, $K_1, \ldots, K_S$ topics are assumed to be mined from the images, respectively. The algorithm inference and learning task of FSSTM for one type of feature are shown in Fig. 6. Given K topics $\beta = (\beta_1, \ldots, \beta_K)$, the log likelihood of $\mathbf{M}$ is defined in

$$\log P(\mathbf{M}) = \sum_{j\in I_{\mathbf{M}}} \mathbf{M}_j \log \sum_{k=1}^{K} \theta_k \beta_{kj} \tag{5}$$

where $I_{\mathbf{M}}$ is the set of term indices of image $\mathbf{M}$, and M_j is the frequency of term j in $\mathbf{M}$. Hence, the inference task is to search for θ_d of the dth image to maximize the likelihood of $\mathbf{M}$. We can obtain

$$\log P(\mathbf{M}) = \sum_{j\in I_{\mathbf{M}}} \mathbf{M}_j \log x_j \tag{6}$$

to set $x_j = \sum_{k=1}^{K} \theta_k \beta_{kj}$ and $x = (x_1, \ldots, x_V)^t$. Differing from the other topic models, FSSTM does not infer θ directly, but reformulates the inference task of optimization over θ as a concave maximization problem over the simplex $\Delta = \mathrm{conv}(\beta_1, \ldots, \beta_K)$ of the topic. It can be seen that x is a convex combination of the K topics $\beta = (\beta_1, \ldots, \beta_K)$ with

$$\sum_k \theta_k = 1, \theta_k \geq 0. \tag{7}$$

FSSTM uses the Frank–Wolfe algorithm [45] as the inference algorithm, which follows the greedy approach. For the lth iteration, we can obtain the topic proportion, as written in

$$\theta_{l+1} := (1 - \alpha')\theta_l \tag{8}$$

where α' and i' are defined by

$$\alpha' := \arg \max_{\alpha\in[0,1]} f(\alpha\beta_{i'} + (1 - \alpha)x_l) \tag{9}$$

$$i' := \arg \max_i \beta_i^t \nabla f(x_l). \tag{10}$$

β_i denotes the standard unit vectors in Δ, and α can be solved by the gradient ascent approach. x_l is a convex combination of at most $L + 1$ vertices of the simplex $\Delta = \mathrm{conv}(\beta_1, \ldots, \beta_K)$ after L iterations, where $x_l = \sum_{k=1}^{K} \theta_{lk}\beta_k$. It implies that at most $l + 1$ out of the K-dimension θ_l are nonzero in FSSTM,

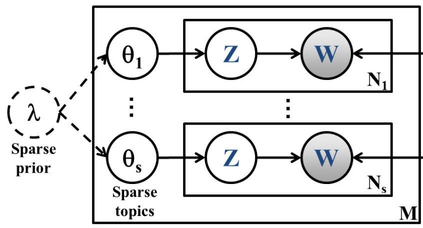

Fig. 7. Probabilistic graphical model of FSSTM.

which can be shown as the implicit constraint $\|\theta\|_0 \leq L+1$. This provides us with a sparse solution, which is obtained by converging at a linear rate to the optimal solutions. By finding $x \in \Delta$ that maximizes the objective function (5), we can infer the latent topic proportion θ_d of the image $\mathbf{M}$.

Given the topic proportion θ_d inferred from the inference task for each image $\mathbf{M}$, the learning task of FSSTM is to learn the topics $\beta = (\beta_1, \ldots, \beta_K)$. The log likelihood of the image data set $\mathbf{C}$ can be expressed as logP($\mathbf{C}$) with the use of Jensens inequality. The lower bound of logP($\mathbf{C}$) with respect to β as written in

$$g(\beta) = \sum_{d \in \mathbf{C}} d_j \sum_{k=1}^{K} \theta_{dk} \log \beta_{kj}, \sum_{j=1}^{V} \beta_{kj} = 1, \quad \beta_{kj} \geq 0 \quad \forall k, j \tag{11}$$

is then maximized using the Lagrangian multipliers. By forcing its derivatives to be zero, the solutions of the topics are updated, as written in

$$\beta_{kj} \propto \sum_{\mathbf{M} \in D} d_j \theta_{dk}. \tag{12}$$

It can be seen that the learning of the topics is simply a multiplication of the new and old representations of the training data. In this way, FSSTM is able to undertake the topic modeling of the image by iterating the inference and learning task until convergence. The optimal sparse solution of latent semantic topic proportion θ is obtained for each type of feature. Then, $\theta_1, \ldots, \theta_S$ are acquired for the representation of each HSR image, all with the implicit prior λ conforming to (2). The probabilistic graphical model of FSSTM is shown in Fig. 7. As can be seen from Fig. 7, when there is one type of feature and S is equal to 1, FSSTM turns out to be FSTM, as shown in Fig. 3. FSSTM designs an adequate feature extraction and fusion strategy to make up for the shortcomings of CAT-FSTM, and is more appropriate for HSR scene classification.

The previous work named SAL-LDA [22] also uses multifeature-based PTM to classify the HSR scenes. SAL-LDA employs spectral, gray-level co-occurrence matrix (GLCM)-based TEX, and SIFT features to describe the image, whereas FSSTM uses wavelets instead of the GLCM as the TEX feature. Wavelets can provide information about both the spatial and frequency contents of an image, which makes them more suitable for analyzing TEX in nonstationary or nonhomogeneous images, e.g., HSR remote sensing images [48]. Differing from SAL-LDA, FSSTM reformulates the inference task as a concave maximization problem, which improves the fusion of the different types of features. The number of topics, which make a nonzero contribution to an image by SAL-LDA, is the same as the topic number, whereas for FSSTM it is less, and does not change much when the topic number is manually changed. This results in FSSTM performing better than SAL-LDA, even with a limited number of training samples. Furthermore, the semantic information generated by SAL-LDA is dense, whereas FSSTM with sparse topic solutions removes the unrepresentative information for the description of the scenes, and the time consumption is greatly reduced.

C. Latent Semantic Fusion and Classification

As mentioned in Section III-A, three features, the spectral, TEX, and structural features, are extracted and clustered to complementarily describe the HSR images, obtaining H_1, H_2, and H_3. Hence, the latent semantics of H_1, H_2, and H_3, denoted as θ_1, θ_2, and θ_3, respectively, are mined by FSSTM. The semantic features θ_1, θ_2, and θ_3 of all the HSR images are then fused at the semantic level, thus obtaining the final semantic feature $F_2 = \{{\theta_1}^T, {\theta_2}^T, {\theta_3}^T\}^T$, with a sparse distribution.

Finally, F_2 with the optimal discriminative characteristics is classified by the SVM classifier with an HIK to predict the scene labels. The HIK measures the degree of similarity between two histograms, to deal with the scale change, and has been applied to image classification using color histogram features [38]. We let $\tilde{\mathbf{V}} = (\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_M)$ be the FSSTM representation vectors of M images, and the HIK is calculated according to

$$K(\tilde{v}_i, \tilde{v}_j) = \sum_k \min(\tilde{v}_{ik}, \tilde{v}_{jk}). \tag{13}$$

In this way, FSSTM provides a complementary feature description, an effective image representation strategy, and an adequate topic modeling procedure for HSR image scene classification, even with limited training samples

IV. Experiments and Analysis

A. Experimental Setup for HSR Scene Classification

The commonly used 21-class UC Merced data set and the 12-class Google data set of the scene image data set designed by the Intelligent Data Extraction and Analysis of Remote Sensing (RS_IDEA) Group in Wuhan University (SIRI-WHU) were used to test the performance of FSSTM. The original large image of the Wuhan IKONOS data set was also used to test the performance of FSSTM in an image annotation application. In the experiments, the images were uniformly sampled with a patch size and spacing of 8 and 4 pixels, respectively. To test the stability of the proposed FSSTM, the different methods were executed 100 times by a random selection of training samples, to obtain convincing results for the three data sets. A k-means clustering operation with the Euclidean distance measurement of the image patches from the training set was employed to construct the visual dictionary, which was the set of V visual words. K topics were selected

TABLE I

OPTIMAL K AND V VALUES FOR THE DIFFERENT METHODS WITH THE UC MERCED DATA SET

Methods	SPE-FSTM	TEX-FSTM	SIFT-FSTM	CAT-FSTM	FSSTM
V	1000	800	1000	2800	2800
K	240	110	250	1000	840

TABLE II

OPTIMAL K AND V VALUES FOR THE DIFFERENT METHODS WITH THE GOOGLE DATA SET OF SIRI-WHU

Methods	SPE-FSTM	TEX-FSTM	SIFT-FSTM	CAT-FSTM	FSSTM
V	1000	800	1000	2800	2800
K	240	300	280	1000	870

TABLE III

OVERALL CLASSIFICATION ACCURACY (%) COMPARISON WITH THE UC MERCED DATA SET

SPM[39]	82.30±1.48
PLSA[18]	89.51±1.31
LDA[19]	81.92±1.12
SAL-LDA[22]	88.33
Cheriyadat[9]	81.67±1.23
Yang and Newsam[8]	81.19
Chen and Tian[40]	89.10
Mekhalfi *et al.*[41]	94.33
Zhao *et al.*[42]	92.92±1.23
Yao *et al.*[21]	93.57±1.02
Hu *et al.*[46]	96.90±0.77
Zhang *et al.*[47]	94.53
SPE-FSTM	78.33±1.42
TEX-FSTM	75.00±1.63
SIFT-FSTM	82.38±1.58
CAT-FSTM	89.76±1.68
FSSTM	**95.71±1.01**

TABLE IV

OVERALL CLASSIFICATION ACCURACY (%) COMPARISON WITH THE GOOGLE DATA SET OF SIRI-WHU

SPM[39]	77.69±1.01
PLSA[18]	89.60±0.89
LDA[19]	60.32±1.20
SAL-LDA[22]	90.65±1.05
Zhao *et al.*[42]	91.52±0.64
SPE-FSTM	83.33±1.06
TEX-FSTM	80.92±0.95
SIFT-FSTM	78.50±1.12
CAT-FSTM	92.83±1.27
FSSTM	**97.83±0.93**

for FSSTM. The visual word number V and topic number K are the two free parameters in the proposed method. Taking the computational complexity and the classification accuracy into consideration, V and K were optimally set, as shown in Tables I and II for the different feature strategies with the UC Merced data set and the Google data set of SIRI-WHU. In Tables I–IV, SPE-FSTM, TEX-FSTM, and SIFT-FSTM denote scene classification utilizing the mean and standard deviation-based spectral feature, the wavelet-based TEX, and the SIFT-based structural feature, respectively.

To further evaluate the performance of FSSTM, the experimental results obtained with the conventional CAT-FSTM, spatial pyramid matching (SPM) [39], PLSA [18], LDA [19], and the semantic allocation level multifeature fusion strategy based on LDA (SAL-LDA) [22] are shown for comparison.

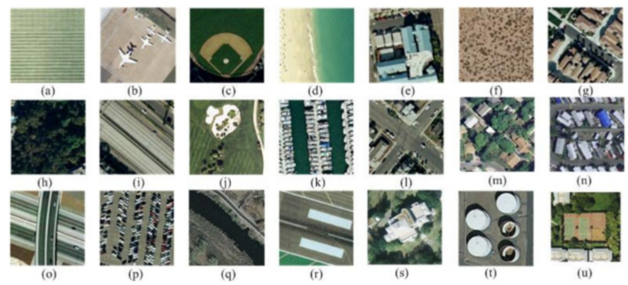

Fig. 8. UC Merced data set. (a) Agricultural. (b) Airplane. (c) Baseball diamond. (d) Beach. (e) Buildings. (f) Chaparral. (g) Dense residential. (h) Forest. (i) Freeway. (j) Golf course. (k) Harbor. (l) Intersection. (m) Medium residential. (n) Mobile home park. (o) Overpass. (p) Parking lot. (q) River. (r) Runway. (s) Sparse residential. (t) Storage tanks. (u) Tennis courts.

We also provide the experimental results obtained with the UC Merced data set, as published in the latest papers by Yang and Newsam [8], Cheriyadat [9], Yao *et al.* [21], Chen and Tian [40], Mekhalfi *et al.* [41], Zhao *et al.* [42], Hu *et al.* [46], and Zhang *et al.* [47]. Dense gray SIFT was employed with SPM, and the spatial pyramid layer was optimally selected as one. In addition, the experimental results obtained for the Google data set of SIRI-WHU with the conventional CAT-FSTM, SPM [39], PLSA [18], LDA [19], and SAL-LDA [22] are shown for comparison, along with the experimental results published in the latest paper by Zhao *et al.* [42].

B. Experiment 1: The UC Merced Image Data Set

The UC Merced data set was downloaded from the USGS National Map Urban Area Imagery collection [8]. This data set consists of 21 land-use scenes (Fig. 8), namely, agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each class contains 100 images and measuring pixels, with a 1-ft spatial resolution. Following the experimental setup published by Yang and Newsam [8], 80 samples were randomly selected per class from the UC Merced data set for training, and the rest were kept for testing.

The classification performance of the single-feature-based FSTM, the conventional multifeature-based FSTM, the proposed FSSTM, and the comparison with the experimental results of previous methods for the UC Merced data set

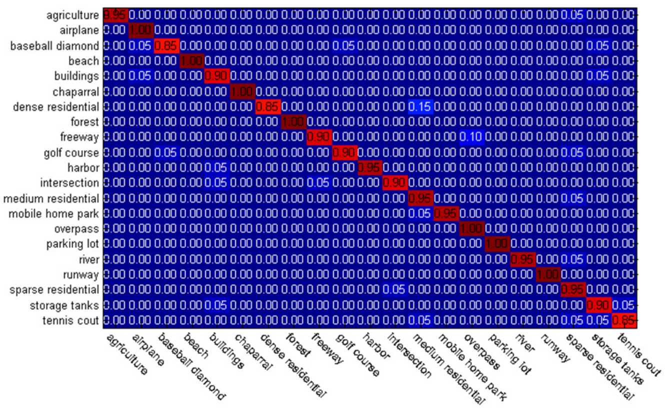

Fig. 9. Confusion matrix of FSSTM with the UC Merced data set.

is reported in Table III. As can be seen in Table III, the classification results of the single-feature-based FSTM and CAT-FSTM are unsatisfactory. The classification accuracy for the proposed FSSTM, $95.71\% \pm 1.01\%$, is the best among all the different methods, and is much better than the result of the single-feature strategy. This indicates that the combination of the semantic feature fusion strategy and sparse representation is able to tradeoff the sparsity and the quality of the inferred semantic information. In addition, it can be seen that FSSTM performs better than SPM [39], PLSA [18], LDA [19], SAL-LDA [22], the Yang and Newsam [8] method, the Cheriyadat [9] method, the Yao *et al.* [21] method, the Chen and Tian [40] method, the Mekhalfi *et al.* [41] method, the Zhao *et al.* [42] method, and the Zhang *et al.* [47] method. Compared with FSSTM, the Hu *et al.* [46] method shows a slight improvement. However, the Hu *et al.* [46] method is implemented by transferring features from pretrained CNNs models, which requires a large number of training samples to train the model.

An overview of the performance of FSSTM is shown in the confusion matrix in Fig. 9. As can be seen in the confusion matrix, most of the scene classes achieve good classification performances, and the airplane, beach, chaparral, forest, overpass, parking lot, and runway scenes can be fully recognized by FSSTM. There is, however, some confusion between certain scenes. For instance, some scenes belonging to the baseball diamond are classified as golf course, storage tanks, and airplane scenes. This may be because all these scenes are composed of a mixture of vegetation cover and bare ground.

To allow a better visual inspection, some of the classification results of CAT-FSTM and FSSTM are shown in Fig. 10.

C. Experiment 2: The Google Data Set of SIRI-WHU

The Google data set of SIRI-WHU[1] was acquired from Google Earth (Google Inc.), covering urban areas in China, and SIRI-WHU [14], [42], [43]. The data set consists of 12 land-use classes, which are labeled as follows: agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water, as shown in Fig. 11. Each class separately contains 200 images, which were cropped to 200×200 pixels, with a spatial resolution of 2 m. In this experiment, 100 training samples were randomly selected per class from the Google data set of SIRI-WHU, and the remaining samples were retained for the testing.

The classification performance of the single-feature-based FSTM, the conventional multifeature-based FSTM, the proposed FSSTM, and the comparison with the experimental results of previous methods for the Google data set of SIRI-WHU is reported in Table IV. As can be seen from Table IV, the classification result of the proposed FSSTM, $97.83\% \pm 0.93\%$, is much better than the spectral, TEX, and SIFT-based FSTM, and the conventional CAT-FSTM method, which confirms that FSSTM, is an effective approach for HSR image scene classification. In Table IV, compared with the other methods, i.e., SPM, SAL-LDA, the LDA method proposed by Lienou *et al.* [19], the PLSA method proposed by Bosch *et al.* [18], and the method of Zhao *et al.* [42], the highest accuracy is acquired by the proposed FSSTM.

[1]The Google data set of SIRI-WHU can be downloaded at http://www.lmars.whu.edu.cn/prof_web/zhongyanfei/e-code.html.

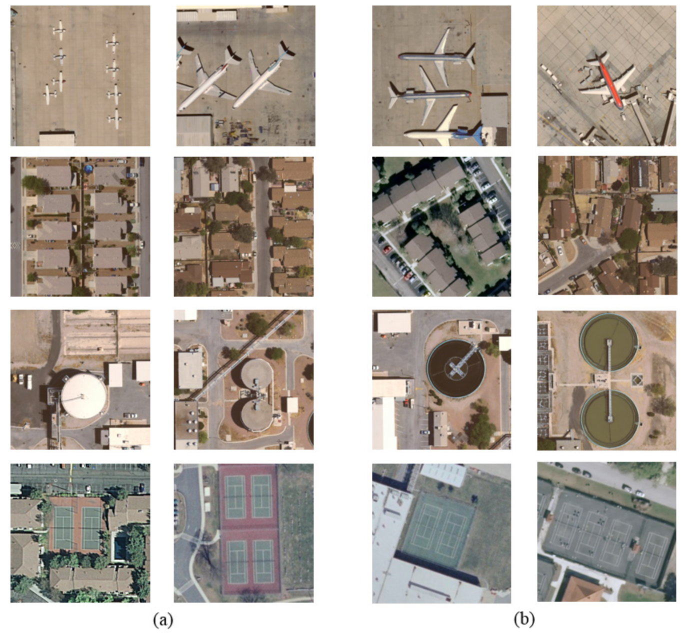

Fig. 10. Some of the classification results of CAT-FSTM and FSSTM. The first, second, third, and fourth lines correspond to the scene classes of airplane, dense residential, storage tanks, and tennis court, respectively. (a) Correctly classified images for all the strategies. (b) Images correctly classified by FSSTM, but incorrectly classified by CAT-FSTM.

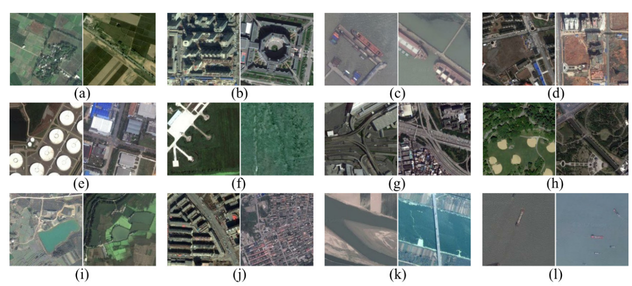

Fig. 11. Google data set of SIRI-WHU. (a) Agriculture. (b) Commercial. (c) Harbor. (d) Idle land. (e) Industrial. (f) Meadow. (g) Overpass. (h) Park. (i) Pond. (j) Residential. (k) River. (l) Water.

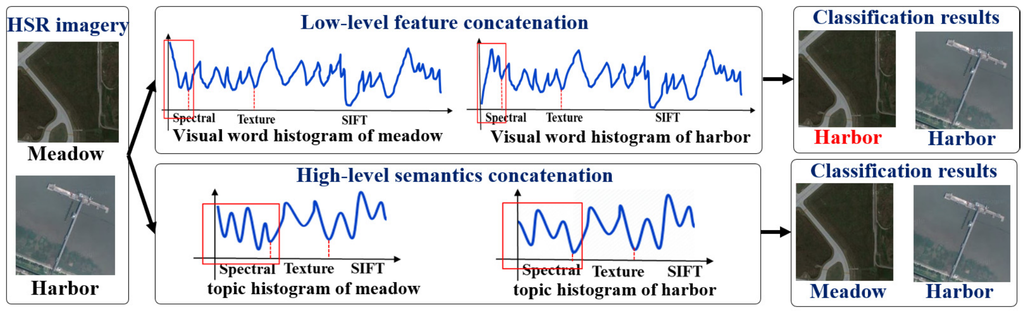

Fig. 12. Examples of scene classification with CAT-FSTM and FSSTM.

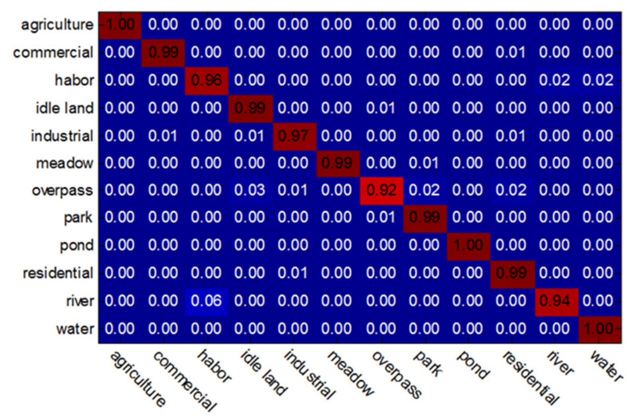

Fig. 13. Confusion matrix of FSSTM with the Google data set of SIRI-WHU.

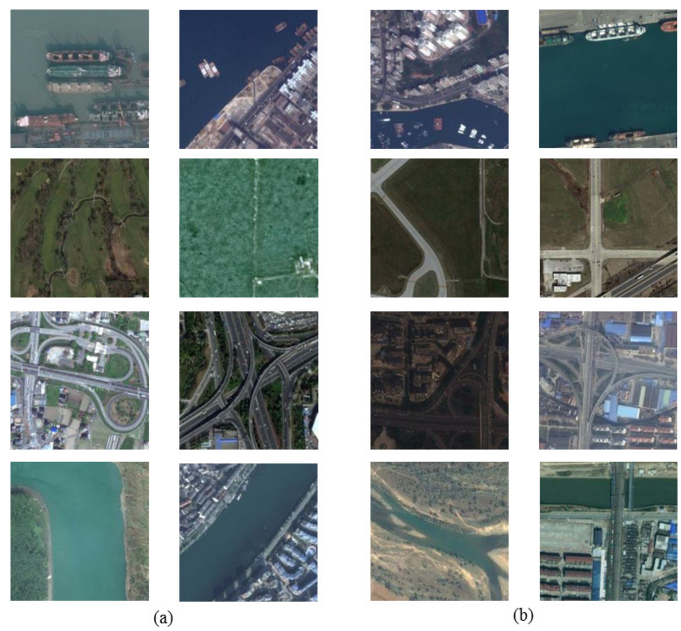

Fig. 14. Some of the classification results of CAT-FSTM and FSSTM. The first, second, third, and fourth lines correspond to the scene classes of harbor, meadow, overpass, and river, respectively. (a) Correctly classified images for all the strategies. (b) Images correctly classified by FSSTM, but incorrectly classified by CAT-FSTM.

To give a clear comparison of FSSTM and the conventional CAT-FSTM, examples of scene classification with CAT-FSTM and FSSTM are shown in Fig. 12, based on the experiments. The spectral, TEX, and SIFT features are extracted from the meadow and harbor scenes, respectively. The TEX and SIFT features of the meadow and harbor scenes are similar, but the spectral features of the two scene images differ a lot. As the proportion of the spectral feature is very small among the three features, by directly concatenating the three types of low-level features and then quantizing the long vectors by clustering, it leads to misclassification for CAT-FSTM. On the other hand, FSSTM fuses the high-level semantics to represent the images, where the proportion of the spectral feature-based topics is the same as the TEX and SIFT features, and the meadow scene can be correctly recognized.

Fig. 13 shows the confusion matrix of FSSTM for the Google data set of SIRI-WHU. On the whole, most of the scene classes achieve good classification performances. There is, however, some confusion between the river scene/harbor scene and the overpass scene/idle land scene. This can be explained by the fact that these classes have similar structural or spectral characteristics, such as both the river scene and harbor scene featuring water and bank.

To allow a better visual inspection, some of the classification results of CAT-FSTM and FSSTM are shown in Fig. 14.

D. Experiment 3: Semantic Annotation of the Wuhan IKONOS Image Data Set

The Wuhan IKONOS data set was acquired by the IKONOS sensor in June 2009, covering the city of Wuhan in China. All of the images in the Wuhan IKONOS data set were obtained by Gram–Schmidt pan-sharpening with ENVI 4.7 software. The spatial resolutions of the panchromatic images and the multispectral images are 1 and 4 m, respectively. The Wuhan IKONOS data set consists of eight land-use scenes, namely, dense residential, idle, industrial, medium residential, parking lot, commercial, vegetation, and water, as shown in Fig. 15. Each class separately contains 30 labeled small images, which were cropped to 150×150 pixels, with a spatial resolution of 1 m. The size of the large image used for the annotation experiment was 6150×8250 pixels, as shown in Fig. 17(a).

In the annotation experiment, the large image was split into a set of small overlapping images of 150×150 pixels. The

TABLE V
ANNOTATION ACCURACIES (%) WITH THE WUHAN IKONOS DATA SET FOR THE DIFFERENT SCENE CLASSIFICATION METHODS

Methods	SPM [39]	PLSA [18]	LDA [19]	CAT-FSTM	Zhao *et al.* [42]	**FSSTM**
Overall accuracy	71.69±7.65	77.34±6.23	84.38±7.24	91.67±2.03	88.96±3.95	**95.83±1.74**

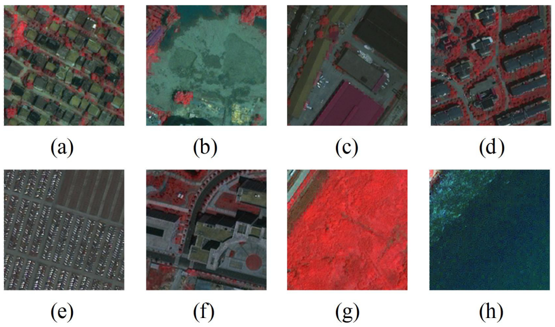

Fig. 15. Training images of the Wuhan IKONOS data set. (a) Dense residential. (b) Idle. (c) Industrial. (d) Medium residential. (e) Parking lot. (f) Commercial. (g) Vegetation. (h) Water.

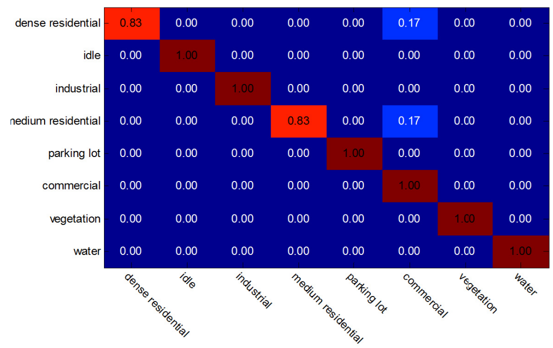

Fig. 16. Confusion matrix of FSSTM with the Wuhan IKONOS data set.

annotation experiment produced good results when the overlap between two adjacent small images was set to 50 pixels. The final label of the overlapping region was decided according to the majority voting method. For the small images, the spectral, TEX, and SIFT features performed well when the patch size and overlap were set to 8×8 pixels and 4 pixels, respectively.

To evaluate the performance of FSSTM, the experimental results obtained with SPM [39], PLSA [18], LDA [19], CAT-FSTM, and the method of Zhao *et al.* [42] are shown for comparison. The different methods were evaluated using the evaluation method published in [19], where 80% of the labeled images were used as training images, and the remaining images were used for testing to evaluate the model. To annotate the large image, all the labeled images were used to train the model. The different methods were executed 20 times by random selection of training samples. To visually evaluate the large annotation maps, the annotation maps were overlaid on the original images with 50% transparency. From Table V, it can be seen that the accuracy of FSSTM, $95.83\% \pm 1.74\%$, is the highest. This confirms the ability of FSSTM to obtain a sparse but representative representation for remote sensing images.

One of the confusion matrices for the Wuhan IKONOS data set was selected from the results obtained by FSSTM, as shown in Fig. 16. From Fig. 16, it can be seen that all of the scenes can be recognized by FSSTM, except for the dense residential and medium building scenes. This is, however, reasonable as the dense residential, the medium residential, and the commercial scenes are composed of similar land-cover objects, such as buildings, trees, roads, and grass. The annotation results obtained with FSSTM for the large Wuhan IKONOS image are shown in Fig. 17(b). As can be seen from Fig. 17(b), there is some confusion between the commercial scene and the medium residential scene, as the buildings in the commercial scene are similar to the buildings in the residential scenes. Due to the fact that the parking lot scene and the industrial scene are similar in the TEX characteristics, the two scene classes also show confusion. Some misclassifications also occur because the large image contains unknown classes and scenes, such as road, school, and gymnasium with special features. Reject classes for areas found in the large image that do not correspond to any of the example images in the training samples have not been defined. Hence, the road scene may be classified as commercial scene, the school scene may be classified as medium residential scene, and the gymnasium scene may be classified as industrial scene. In addition, there are obvious patch effects on the edge of the diverse scenes in the annotated image. This is a result of the sampling method. In order to avoid such problems, we plan in our future work to combine the annotation with geographic data. However, from the perspective of our remote sensing image analysis expertise, the overall annotation performance is still satisfactory.

V. SENSITIVITY ANALYSIS

A. Sensitivity Analysis in Relation to the Topic Number K

To investigate the sensitivity of CAT-FSTM and FSSTM in relation to the topic number K, the values of the patch size, the patch spacing, and the visual word number V were kept constant at 8, 4, and 2800, respectively. The topic number K was then varied over the range of [400, 600, 800, 1000, 1200] for the UC Merced data set and Google data set of SIRI-WHU. As shown in Fig. 18, with the increase of the topic number K, the overall accuracies (OAs) of FSSTM and CAT-FSTM both become higher and then tend to decline. FSSTM obtains the highest accuracy when K is 800, while CAT-FSTM demands more than 1000 topics. This indicates that FSSTM can obtain a better performance with fewer topics.

B. Sensitivity Analysis in Relation to the Visual Word Number V

To study the sensitivity of CAT-FSTM and FSSTM in relation to the visual word number V, the values of the

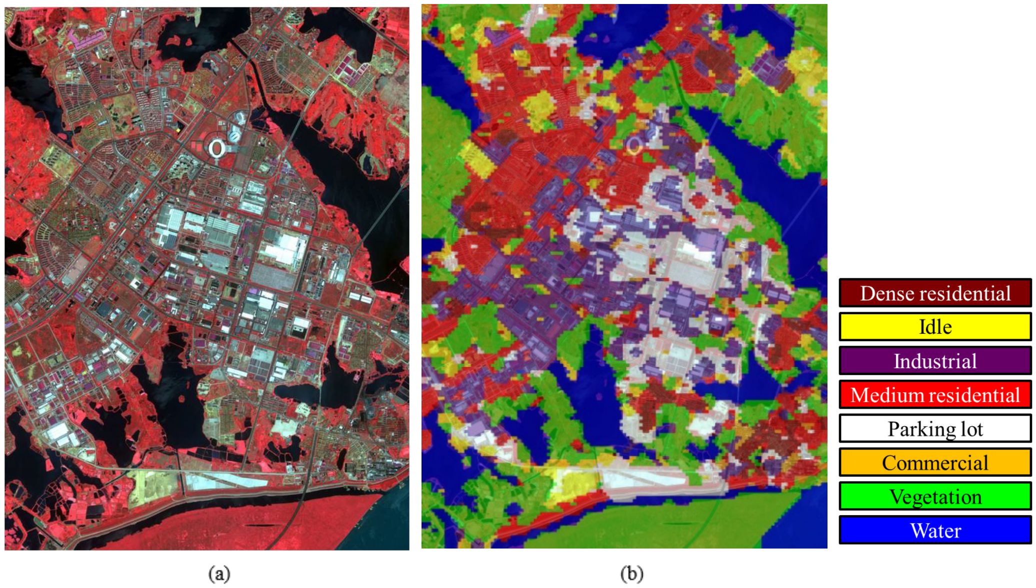

Fig. 17. Annotated image obtained by FSSTM. (a) False-color image to be annotated. (b) Annotated image.

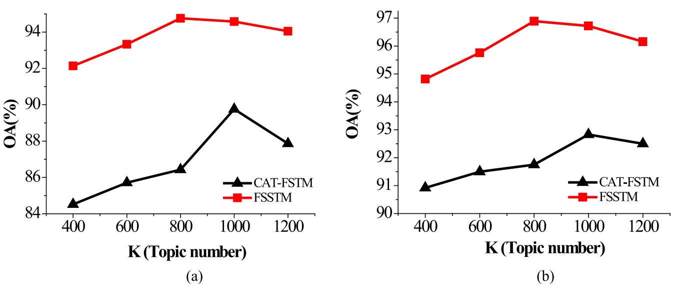

Fig. 18. Sensitivity analysis of CAT-FSTM and FSSTM in relation to the topic number K. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

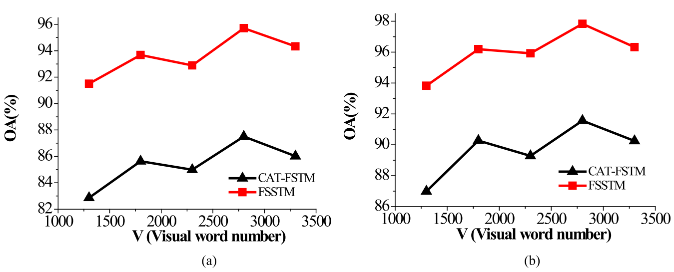

Fig. 19. Sensitivity analysis of CAT-FSTM and FSSTM in relation to the visual word number V. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

patch size, the patch spacing, and the topic number K were kept constant at 8, 4, and 840, respectively. The visual word number V was then varied over the range of [1300, 1800, 2300, 2800, 3300] for the UC Merced data set and the Google data set of SIRI-WHU. Comparing Figs. 18 and 19, it can be seen that the OA curves of FSSTM and CAT-FSTM display a greater fluctuation in relation to the visual word number V, and they are less sensitive to the topic number K. It is notable that FSSTM is superior to CAT-FSTM over the entire range for the two data sets, which infers that the proposed FSSTM can outperform the traditional fusion strategy.

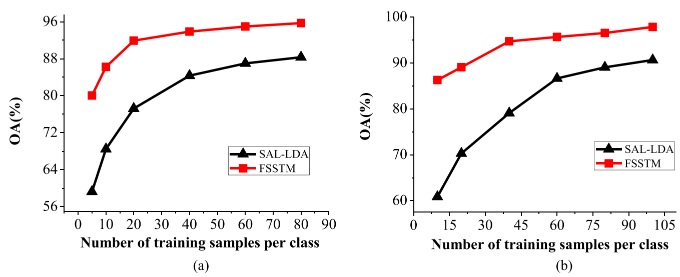

Fig. 20. Classification accuracies with different numbers of training samples per class. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

C. Sensitivity Analysis in Relation to the Number of Training Samples

By modeling the large collection of images with only a few latent topic proportions of nonzero values, we addressed the situation of HSR imagery with limited training samples, employing FSSTM and SAL-LDA [22], respectively. The training number was varied over the range of [80, 60, 40, 20, 10, 5] for the UC Merced data set, and the training number for the Google data set of SIRI-WHU was varied over the range of [100, 80, 60, 40, 20, 10]. The OAs obtained with different numbers of training samples for the UC Merced data set and the Google data set of SIRI-WHU are reported in Fig. 20.

As can be seen from Fig. 20, when compared with SAL-LDA, the proposed FSSTM performs the best, and is relatively stable with the decrease in the number of training samples per class for the two data sets. When the training samples are under 20%, or even 10% or 5%, FSSTM displays a smaller fluctuation than SAL-LDA, and can obtain a satisfactory and robust performance with limited training samples.

D. Sensitivity Analysis of the Topic Modeling Sparsity in Relation to the Topic Number K

In order to investigate the sparsity of the latent semantics for FSSTM and SAL-LDA in relation to the topic number K, the

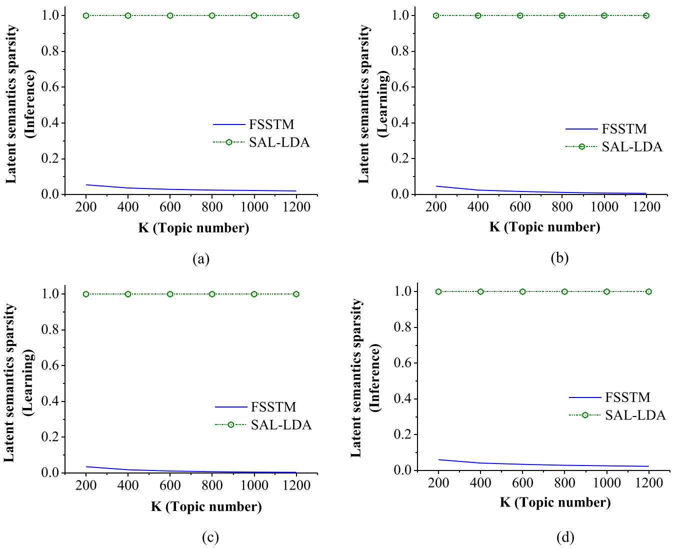

Fig. 21. Sensitivity analysis of the latent semantic sparsity for FSSTM and SAL-LDA in relation to the topic number K. (a) and (b) UC Merced data set. (c) and (d) Google data set of SIRI-WHU.

values of the patch size, the patch spacing, and the visual word number V were kept constant at 8, 4, and 2800, respectively. The topic number K was then varied over the range of [200, 400, 600, 800, 1000, 1200] for the UC Merced data set and Google data set of SIRI-WHU. The latent semantic sparsity is used to determine the sparsity level of the latent semantics mined by the topic model when conducting the inference and learning task. From Fig. 21, it can be seen that SAL-LDA is unable to discover the sparse latent semantics in both the inference and learning phases for the two data sets. In contrast, FSSTM is able to discover very sparse latent semantics in both the inference and learning phases. The latent semantic sparsity of FSSTM decreases with the increase of the topic number K. For example, when 1000 topics are set to be mined, only about 22 topics make a nonzero contribution to the HSR image, and about 20 topics make a nonzero contribution to the image when 800 topics are set. This is consistent with the fact that a specific image usually contains only a few specific topics, which contribute a lot to the semantic understanding of the image. Accordingly, FSSTM can robustly model the HSR images with sparse latent semantics, as well as adequate latent semantics.

E. Sensitivity Analysis of the Topic Modeling Time in Relation to the Topic Number K

To investigate the time efficiency of the proposed FSSTM and the conventional CAT-FSTM and SAL-LDA, the values of the patch size, the patch spacing, and the visual word number V were kept at the optimal parameter settings, respectively. The topic number K was then varied over the range of [100, 300, 500, 700, 900] for the UC Merced data set and the Google data set of SIRI-WHU. As can be seen from Fig. 22, the topic modeling time of SAL-LDA far transcends the modeling time of CAT-FSTM and FSTM. With the increase of the topic number K, the time curve of SAL-LDA displays

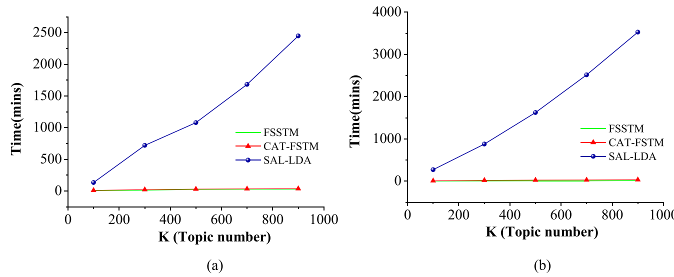

Fig. 22. Time efficiency of the different methods in relation to the topic number K. (a) UC Merced data set. (b) Google data set of SIRI-WHU.

TABLE VI
TIME EFFICIENCY (MINUTES) OF THE DIFFERENT METHODS IN RELATION TO THE TOPIC NUMBER K WITH THE UC MERCED DATA SET

K	SAL-LDA	CAT-FSTM	FSSTM
100	135	12	6
300	720	23	14
500	1080	31	26
700	1683	35	30
900	2448	39	31

TABLE VII
TIME EFFICIENCY (MINUTES) OF THE DIFFERENT METHODS IN RELATION TO THE TOPIC NUMBER K WITH THE GOOGLE DATA SET OF SIRI-WHU

K	SAL-LDA	CAT-FSTM	FSSTM
100	270	9	3
300	879	21	5
500	1623	25	8
700	2517	28	9
900	3528	33	12

linear growth, and the time curves of SAL-FSTM and FSSTM stay relatively smooth. From Tables VI and VII, it can be seen that the modeling time of FSSTM is the shortest among the three methods. For example, even when the topic number is 900 for the Google data set of SIRI-WHU, SAL-LDA requires 3528 min to discover the semantics, CAT-FSTM requires 33 min, while FSSTM needs only 12 min to discover the semantics for the data set. This indicates that FSSTM is an efficient PTM compared with the classical nonsparse PTMs, such as SAL-LDA, and the conventional sparse PTM.

VI. CONCLUSION

In this paper, we have designed an effective approach the FSSTM for HSR imagery scene classification. FSSTM is proposed to address the scene classification and annotation problem in sparse topic modeling, and is able to ensure that the discovered semantics are both sparse and adequate. The proposed approach fuses the distinct sparse semantics at the semantic level and makes full use of the complex characteristics of HSR images. As a robust approach, FSSTM obtains satisfactory performances, even with a limited number of training samples. Evaluations of the classification and annotation experiments undertaken in this paper showed that the proposed

FSSTM method can discover sparser and more discriminative semantics in a shorter time than the conventional PTM.

Nevertheless, image patches obtained by the uniform grid sampling method may be unable to preserve the semantic information of a complete scene. It would, therefore, be desirable to combine image segmentation with scene classification. The clustering strategy, as one of the most important techniques in remote sensing image processing, is another point that should be considered. HSR images with different resolutions from diverse remote sensing sensors are often processed together in practical use. As future extensions, we plan to consider data sets acquired from different sensor types or with different resolutions. The design of appropriate features is also a significant problem for HSR image scene classification. Hence, in our future work, we plan to explore more representative features, based on the analysis of multi-source data sets.

REFERENCES

[1] R. Bellens, S. Gautama, L. Martinez-Fonte, W. Philips, J. C. W. Chan, and F. Canters, "Improved classification of VHR images of urban areas using directional morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 2803–2813, Oct. 2008.

[2] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.

[3] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, and J. C. Tilton, "Learning Bayesian classifiers for scene classification with a visual grammar," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 581–589, Mar. 2005.

[4] R. Bahmanyar, S. Cui, and M. Datcu, "A comparative study of bag-of-words and bag-of-topics models of EO image patches," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1357–1361, Jun. 2015.

[5] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[6] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2015, pp. 44–51.

[7] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, Jun. 2016.

[8] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Int. Conf. ACM SIGSPATIAL GIS*, San Jose, CA, USA, 2010, pp. 270–279.

[9] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[10] B. Zhao, Y. Zhong, and L. Zhang, "Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery," *Remote Sens. Lett.*, vol. 4, no. 12, pp. 1204–1213, Dec. 2013.

[11] N. Serrano, A. E. Savakis, and J. Luo, "Improved scene classification using efficient low-level features and semantic cues," *Pattern Recognit.*, vol. 37, no. 9, pp. 1773–1784, 2004.

[12] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 2. Oct. 2003, pp. 1470–1477.

[13] L.-J. Zhao, P. Tang, and L.-Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.

[14] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[15] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, Jan./Feb. 2001.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 5, pp. 993–1022, Mar. 2003.

[17] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[18] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.

[19] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.

[20] C. Vaduva, I. Gavat, and M. Datcu, "Latent Dirichlet allocation for spatial analysis of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2770–2786, May 2013.

[21] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[22] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[23] W. Shao, W. Yang, G.-S. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Computer Vision Systems*. Berlin, Germany: Springer, 2013, pp. 324–333.

[24] X. Zheng, X. Sun, K. Fu, and H. Wang, "Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 652–656, Jul. 2013.

[25] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Trans. Geosci Remote Sens.*, vol. 53, no. 1, pp. 280–295, Jan. 2015.

[26] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1313–1320.

[27] J. Zhu and E. P. Xing, "Sparse topical coding," in *Proc. Uncertainty Artif. Intell. (UAI)* 2011, pp. 831–838.

[28] K. Than and T. B. Ho, "Fully sparse topic models," in *Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Discovery Databases (ECML PKDD)*, vol. 7523. Bristol, U.K., 2012, pp. 490–505.

[29] R. Fernandez-Beltran and F. Pla, "Incremental probabilistic latent semantic analysis for video retrieval," *Image Vis. Comput.*, vol. 38, pp. 1–12, Jun. 2015.

[30] R. Kaviani, P. Ahmadi, and I. Gholampour, "Incorporating fully sparse topic models for abnormality detection in traffic videos," in *Proc. IEEE Conf. Comput. Knowl. Eng.*, Oct. 2014, pp. 586–591.

[31] W. Yang, X. Yin, and G. S. Xia, "Learning high-level features for satellite image classification with limited labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4472–4482, Aug. 2015.

[32] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[33] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 193–204, Mar. 2011.

[34] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[35] K. Huang and S. Aviyente, "Wavelet feature selection for image classification," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1709–1720, Sep. 2008.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jan. 2004.

[37] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 524–531.

[38] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2003, pp. III-513–III-516.

[39] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2169–2178.

[40] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.
[41] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "Land-use classification with compressive sensing multifeature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2155–2159, Oct. 2015.
[42] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.
[43] B. Zhao, Y. Zhong, L. Zhang, and B. Huang, "The Fisher kernel coding framework for high spatial resolution scene classification," *Remote Sens.*, vol. 8, no. 2, p. 157, 2016.
[44] Y. Zhong, M. Cui, Q. Zhu, and L. Zhang, "Scene classification based on multifeature probabilistic latent semantic analysis for high spatial resolution remote sensing images," *J. Appl. Remote Sens.*, vol. 9, no. 1, p. 095064, 2015.
[45] K. L. Clarkson, "Coresets, sparse greedy approximation, and the Frank–Wolfe algorithm," *ACM Trans. Algorithms*, vol. 6, no. 4, pp. 63:1–63:30, 2010.
[46] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.
[47] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.
[48] S. D. Newsam and C. Kamath, "Retrieval using texture features in high-resolution multispectral satellite imagery," *Proc. SPIE*, vol. 5433, pp. 21–32, Apr. 2004.

Qiqi Zhu received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2013. She is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

Her research interests include scene analysis for high spatial resolution remote sensing imagery and topic modeling.

Yanfei Zhong (M'11–SM'15) received the B.S. degree in information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He has been a Teacher of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, since 2007, and was promoted to an Associate Professor and a Full Professor in 2008 and 2010, respectively. He has authored over 120 research papers, including over 60 peer-reviewed articles in international journals, such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, and *Pattern Recognition*. His research interests include hyperspectral remote sensing information processing, high-resolution remote sensing image understanding, and geoscience interpretation for multisource remote sensing data.

Dr. Zhong was a recipient of the Excellent Young Scientist Foundation (selected by the National Natural Science Foundation of China), the National Excellent Doctoral Dissertation Award of China, and New Century Excellent Talents in University of China (selected by the Ministry of Education of China). He was also a Referee of over 30 international journals. He is serving as an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, *International Journal of Remote Sensing*, and *Remote Sensing*.

Liangpei Zhang (M'06–SM'08) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xian Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently the Head of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, also a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China, and also a Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He has over 450 research papers and five books. He holds 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a Fellow of the Institution of Engineering and Technology. He was a recipient of the 2010 Best Paper Boeing Award and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing. He regularly serves as a Co-Chair of the series SPIE conferences on multispectral image processing and pattern recognition, conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor of the *International Journal of Ambient Computing and Intelligence*, the *International Journal of Image and Graphics*, the *International Journal of Digital Multimedia Broadcasting*, the *Journal of Geospatial Information Science*, and the *Journal of Remote Sensing*, and the Guest Editor of the *Journal of Applied Remote Sensing* and the *Journal of Sensors*. He is currently serving as an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Deren Li (M'02–SM'03) received the M.Sc. degree in photogrammetry and remote sensing from the Wuhan Technical University of Surveying and Mapping, Wuhan University, Wuhan, China, in 1981, and the Doctor of Engineering degree in photogrammetry and remote sensing from Stuttgart University, Stuttgart, Germany, in 1985.

He was elected as an Academician of the Chinese Academy of Sciences in 1991, and CAE and Euro-Asia Academy of Sciences in 1995. He is currently the Academic Committee Chairman of the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include spatial information science and technology, such as remote sensing, Global Positioning System, and Geographic Information System, and their integration.

Dr. Li was the President of the ISPRS Commissions III and VI and the first President of the Asia GIS Association from 2002 to 2006.