

Scene Classification Based on the Multifeature Fusion Probabilistic Topic Model for High Spatial Resolution Remote Sensing Imagery

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

Abstract—Scene classification has been proved to be an effective method for high spatial resolution (HSR) remote sensing image semantic interpretation. The probabilistic topic model (PTM) has been successfully applied to natural scenes by utilizing a single feature (e.g., the spectral feature); however, it is inadequate for HSR images due to the complex structure of the land-cover classes. Although several studies have investigated techniques that combine multiple features, the different features are usually quantized after simple concatenation (CAT-PTM). Unfortunately, due to the inadequate fusion capacity of k -means clustering, the words of the visual dictionary obtained by CAT-PTM are highly correlated. In this paper, a semantic allocation level (SAL) multifeature fusion strategy based on PTM, namely, SAL-PTM (SAL-pLSA and SAL-LDA) for HSR imagery is proposed. In SAL-PTM: 1) the complementary spectral, texture, and scale-invariant-feature-transform features are effectively combined; 2) the three features are extracted and quantized separately by k -means clustering, which can provide appropriate low-level feature descriptions for the semantic representations; and 3) the latent semantic allocations of the three features are captured separately by PTM, which follows the core idea of PTM-based scene classification. The probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) models were compared to test the effect of different PTMs for HSR imagery. A U.S. Geological Survey data set and the UC Merced data set were utilized to evaluate SAL-PTM in comparison with the conventional methods. The experimental results confirmed that SAL-PTM is superior to the single-feature methods and CAT-PTM in the scene classification of HSR imagery.

Index Terms—Fusion, high spatial resolution (HSR) imagery, latent Dirichlet allocation (LDA), multifeature, probabilistic latent semantic analysis (pLSA), probabilistic topic model (PTM), scene classification.

I. INTRODUCTION

WITH the ongoing development of satellite sensors, huge quantities of high spatial resolution (HSR) remote sensing images have now become available. Nevertheless, this type

of data demonstrates the phenomena of a complex spatial arrangement with high intraclass and low interclass variabilities, which poses a big challenge for image classification. According to these characteristics, the classification methods for HSR images have evolved from per-pixel-oriented methods to object-oriented methods. Both object-based and contextual-based methods can achieve precise object recognition [1]–[5]. However, these methods have no access to the semantics in the image. This leads to the so-called “semantic gap,” namely, the divergence between the low-level data and the high-level semantic information [6]. In order to acquire the semantic information in accordance with human cognition, how to effectively utilize the strengths of the HSR images is a big issue. Therefore, scene classification, which is aimed at automatically labeling an image from a set of semantic categories [7], has been proposed and has shown remarkable success in image interpretation. To date, it has been systematically studied in natural image analysis [8]–[10]. For HSR image analysis, scene representation and recognition is a challenging task owing to the ambiguity and variability of the scenes. For example, given a set of HSR images containing different scenes, the land-cover objects can be recognized based on the low-level feature description, e.g., buildings. However, the capture of high-level latent semantic concepts, such as residential, commercial, and industrial areas, which usually contain a variety of land-cover objects, is a challenging problem. In other words, the main problem in HSR image semantic interpretation is to bridge the semantic gap [34]. As a consequence, scene classification, as an effective means of HSR image semantic interpretation, has been widely applied [11]–[13]. For instance, in [14], Cheriadat explored an unsupervised feature learning approach to directly model an aerial scene by exploiting the local spatial and structural patterns for scene classification. Yang *et al.* [49] proposed spatial pyramid co-occurrence, which can represent land-use scenes from both the photometric and geometric aspects. Furthermore, in [15], Aksoy *et al.* assigned satellite images to different scene classes under a Bayesian framework.

Among the scene classification methods, object-based scene classification [15]–[17] utilizes a relevant model to define the spatial relationship between the objects, based on the recognition of objects (such as roads, trees, and grass). For this approach, prior information about the objects is required. The object recognition needs to be well designed, and the spatial relationship is difficult to model. Therefore, the bag of words

Manuscript received September 16, 2014; revised January 20, 2015 and March 27, 2015; accepted May 13, 2015. This work was supported by the National Natural Science Foundation of China under Grant 41371344, and the Fundamental Research Funds for the Central Universities under Grant No. 2042014kf00231. (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: zhongyanfei@whu.edu.cn; wuxi5477@126.com; zlp62@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.2015.2435801

(BOW) model-based approach for scene classification has been receiving more and more attention [18]–[22]. Inspired by the BOW model assumption [23], the probabilistic topic model (PTM), including probabilistic latent semantic analysis (pLSA) [24] and latent Dirichlet allocation (LDA) [25], represents the imagery as a random mixture of topics. The PTM has been widely applied in the natural image scene analysis field [26]–[32]. In recent years, more and more researchers have employed the PTM to solve the challenges of HSR image scene classification [33]. In general, a single feature, e.g., the color, structural, shape, or texture feature, is utilized to describe the visual words. For instance, Liénou *et al.* [34] employed the spectral feature (mean and standard deviation) as the feature descriptor for the patches. Xu *et al.* [35] used scale-invariant feature transform (SIFT) as the feature extractor. However, due to the complex structure and abundant information in HSR images, it is widely accepted that multiple features—such as features based on texture, color, and structural information—should be adaptively fused to discriminate each class from the others. Wang *et al.* [36] combined visual, object, and spatial relationship semantic features for image retrieval. Sheng *et al.* [13] designed sparse coding-based multiple feature fusion (SCMF) for HSR image scene classification. SCMF sets the fused result as the concatenation of the probability images obtained by the sparse codes of SIFT, the local ternary pattern histogram Fourier, and the color histogram features. Shao *et al.* [11] fused the probability output of SIFT, tree of colored shapes (tree of c-shapes), the discriminative completed local binary pattern, and the bag-of-colors features to characterize HSR images. Zheng *et al.* [37] also used four features and concatenated the quantized vector by k -means clustering for each feature to form a vocabulary. Faria *et al.* [38] proposed a framework for classifier fusion and for selecting the most appropriate classifier, based on diversity image descriptors and learning methods. The experiments undertaken by Faria *et al.* showed that the framework could achieve results that were comparable to some well-known algorithms in the literature. In addition, some studies have employed a boosting classifier to choose optimal features for the training set [39], [40].

Methods combining multiple features have also been proposed for PTM-based satellite image scene classification. Luo *et al.* [41], [42] constructed descriptors for a satellite image by concatenating the color and texture features and then quantized the descriptors of all the patches into several kinds of visual words by k -means clustering. However, different features lead to different feature descriptors, and they usually differ greatly. For instance, the spectral feature value can reach 1–255, but the SIFT feature value is usually very small, even 10^{-20} . When k -means clustering is used to quantize the vector concatenated by multiple feature descriptors, such as the spectral, texture, and structural features, the different features interfere with each other. Furthermore, k -means clustering is inadequate to capture the abundant spectral information and complex structure in HSR images. This leads to a visual dictionary in which the visual words are highly correlated. Hence, to unify the order of magnitude, the feature values are usually normalized before fusion. However, due to the inadequate fusion capacity of k -means clustering and the mutual interference between different features, the normalization operation contributes little to conserv-

ing and exploiting the useful information. We call this method “visual word level multifeature concatenation” (CAT-PTM).

Inspired by the aforementioned work, we present a semantic allocation level (SAL) multifeature fusion strategy based on PTM, namely, SAL-PTM (SAL-pLSA and SAL-LDA), for HSR imagery. The main contributions of this paper are as follows.

1) *Effective Feature Description Method for HSR Imagery:* Considering the distinct characteristics of HSR imagery, we choose the mean and standard deviation for the spectral feature, the gray-level co-occurrence matrix (GLCM) [43] for the texture feature, and SIFT [44] for the structural feature in SAL-PTM. The three features are combined to exploit the spectral and spatial information of the HSR imagery. An HSR image can then be described by the three different feature vectors.

2) *Appropriate Image Representation Generation Strategy for HSR Imagery:* The spectral, texture, and SIFT feature vectors are quantized separately to generate three 1-D histograms during the k -means clustering. In this way, the visual dictionary generated by the statistical frequency of the histograms for all the images contains visual words that are uncorrelated and is an appropriate image representation for the topic model.

3) *Adequate Latent Semantic Allocation Mining Procedure for PTM-Based HSR Image Scene Classification:* As for PTM-based HSR image scene classification, the core idea is to automatically capture the most discriminative latent semantic allocations. Hence, the latent semantic allocations of the three image representations are captured separately by PTM and are then fused into the final latent semantic allocation vector. This semantic mining procedure retains the distinctive characteristics of each image representation. Finally, a support vector machine (SVM) with the radial basis function (RBF) kernel is employed to predict the scene labels.

The proposed SAL-PTM was evaluated and compared to the conventional CAT-PTM and single-feature methods. Four-class U.S. Geological Survey (USGS) aerial orthophotographs, the 21-class UC Merced data set, and an original large image from the four-class USGS data set were applied here for the testing. The experimental results confirmed that the proposed method is superior to CAT-PTM and the single-feature methods.

The rest of this paper is organized as follows. In Section II, we describe the PTM, including pLSA and LDA. Section III provides details about the proposed SAL-PTM for HSR imagery scene classification. A description of the data sets and an analysis of the experimental results are presented in Section IV. Section V provides the result of the sensitivity analysis. Finally, the conclusions are drawn in Section VI.

II. BACKGROUND

The PTM, which includes pLSA and LDA, introduces a latent variable to analyze the visual words generated by the BOW model. First proposed in natural language processing, BOW has been widely applied in image interpretation due to the similarity between text analysis and image processing. Given a data set consisting of M images, each image can be described by a set of N visual words w_n from a visual dictionary. The data set can then be represented as a word-image co-occurrence matrix, where each element denotes the occurrence number of a visual word in an image.

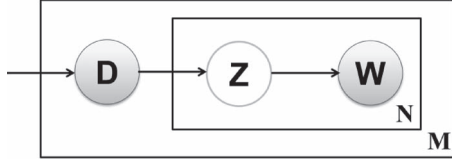


Fig. 1. Probabilistic graphical model of pLSA. The nodes D, Z, and W represent image, topic, and visual word, respectively.

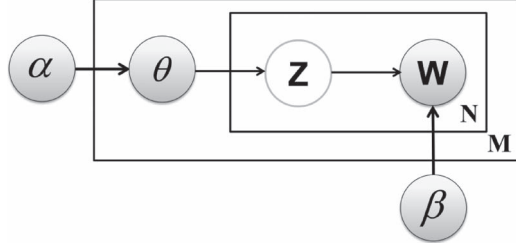


Fig. 2. Probabilistic graphical model of LDA.

pLSA was first proposed by Hofmann in 1990 [24], and it utilizes a graphical model to represent the relationship between image, topic, and visual word, as shown in Fig. 1.

By combining probability and statistics theory with the BOW representation, each element in the co-occurrence matrix can be transformed into the joint probability $p(w_j, d_i)$, which denotes the probability of the visual word appearing in image d_i . On the basis of a conditional probability formula, we decompose the $V \times M$ probability co-occurrence matrix

$$p(w_j, d_i) = p(d_i)p(w_j|d_i). \quad (1)$$

In a similar way, we decompose $p(w_j, d_i)$ on the basis of the total probability formula

$$p(w_j|d_i) = \sum_{k=1}^K p(w_j|z_k)p(z_k|d_i). \quad (2)$$

In (2), $\{p(w_j|z_1), \dots, p(w_j|z_k), \dots, p(w_j|z_K), w_j \in W\}$ forms a set of base vectors which expand into a latent semantic space, and the mixing weights $p(z_k|d_i)$ denote the image specific topic probability distribution, namely, the latent semantics that we intend to mine. Therefore, with the introduction of the pLSA model, we represent each image as a set of vectors $\{p(z_1|d_i), \dots, p(z_k|d_i), \dots, p(z_K|d_i), d_i \in D\}$, which is the input of the classifier.

In the pLSA model, it can be seen that each image is merely the hybrid digital expression of the discrete probabilities of the topics, which leads to the overfitting phenomenon. In addition, the pLSA model is unable to assign probabilities to images outside of the training samples.

In order to overcome the shortcomings of pLSA, in 2003, Blei [25] proposed LDA. Based on pLSA, LDA treats the topic mixture parameters as variables drawn from a Dirichlet distribution; that is, for an image data set, given K topics, the K -dimensional random variable $\theta = \{\theta_1, \dots, \theta_i, \dots, \theta_M\}$, where $\theta_i = \{\theta_{i1}, \dots, \theta_{ik}, \dots, \theta_{iK}\}$ follows a Dirichlet distribution, whose parameter is $\alpha = \{\alpha_1, \dots, \alpha_i, \dots, \alpha_K\}$. The LDA model is represented as a graphical model in Fig. 2. LDA defines a probability function for the original discrete latent semantic distribution, making up for the shortcomings of pLSA.

III. SCENE CLASSIFICATION BASED ON THE MULTIFEATURE FUSION PTM FOR HSR REMOTE SENSING IMAGERY

In the proposed SAL-PTM, we employ the LDA model and the pLSA model to capture the semantic information from the HSR images, on the basis of an appropriate image representation generation strategy and an adequate latent semantic allocation mining procedure.

The previous studies have shown that an even grid sampling strategy yields a better classification performance than other sampling strategies such as random sampling [29]. Hence, all the images are split into image patches using an even grid sampling strategy. The image patches are digitized by the spectral, texture, and SIFT features, respectively, thus making up three sets of feature vectors: the spectral feature vector (Spe), the texture feature vector (Tex), and the SIFT feature vector (Sif). However, with the influence of illumination, rotation, and scale variation, the same visual word in different images may be endowed with various feature values. A k -means clustering operation is applied to generate 1-D frequency vector histograms. In this way, image patches with similar feature values can correspond to the same visual word. A statistical analysis of the frequency for each visual word is performed for all the images. Three word-image co-occurrence matrices, Mat_{spe} , Mat_{tex} , and Mat_{sif} , are thus acquired. Each column of the matrices represents an image, and each element denotes the frequency of the visual word. The co-occurrence matrix generation process of an image is shown in Fig. 3. Then the multifeature latent semantic allocation vector is mined by SAL-PTM.

The next procedure is to input the vector into a proper classifier to predict the scene labels. In this paper, SVM is utilized to test the ability of the discrimination for the latent semantic allocation. The core idea of SVM [45] is to effectively train a linear learning classifier in the kernel function space, which can solve the pattern classification problem in a nonlinear way, as well as give consideration to the generalization and optimization performance. SVM is built based on the structural-risk-minimization principle and Vapnik–Chervonenkis (VC) dimension theory, in which kernel functions and parameters are chosen. In this way, a bound on the VC dimension is minimized [46]. Several researchers have employed SVM as the classifier for scene classification [28], [47], [48]. Bosch *et al.* [28] utilized SVM with the RBF kernel as the scene classifier. The RBF kernel is able to handle the case where the relationship between the class labels and attributes is nonlinear [49]. It is an appropriate choice for the kernel function of SVM and is therefore applied in this paper. The overall procedure for scene classification based on the multifeature fusion PTM is shown in Fig. 4.

A. Multiple Feature Combination

Diverse features should be combined, and the reasons for the inadequacy of employing a single feature are manifold. HSR images have abundant spectral information and are also rich in spatial information. Among the feature descriptors, the spectral feature descriptor is the reflection of the attributes that constitute the ground components. The texture feature descriptor contains information about the spatial distribution of tonal

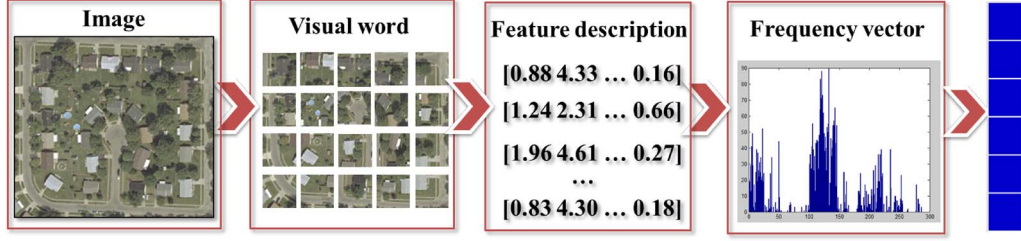


Fig. 3. Co-occurrence matrix generation process of an image, where the column of blue blocks represents the word-image co-occurrence matrix of the image.

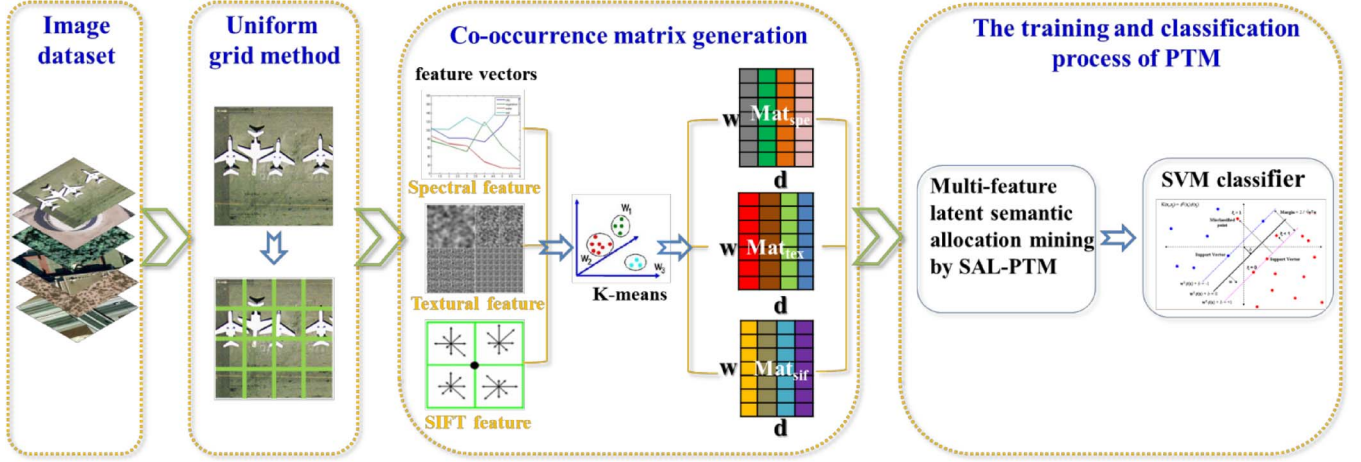


Fig. 4. Flowchart of scene classification based on the multifeature PTM for HSR imagery.

variations within a band [50]. The SIFT feature descriptor can overcome affine transformations, changes in the illumination, and changes in the 3-D viewpoint and has thus been widely applied in image analysis [14], [44], [51]. A comprehensive description of the HSR images depends upon the specific combination of multiple complementary features, such as the spectral feature, texture feature, and SIFT feature. The method proposed in this paper utilizes the three features to fully depict the visual words and provides abundant descriptions for the follow-up scene classification work.

- 1) For the spectral feature, the values of the mean and standard deviation for each visual word of each band are calculated according to (3) and (4), respectively

$$\text{mean} = \frac{\sum_{i=1}^n v_i}{n} \quad (3)$$

$$\text{std} = \sqrt{\frac{\sum_{i=1}^n (v_i - \text{mean})^2}{n}} \quad (4)$$

In (3) and (4), n is the total number of image pixels, and v_i denotes the i th pixel gray value of the image band. We assume that there are B bands, and the k th band conforms to $k \in (1, B)$. It is noted that mea_k denotes the mean value of the k th band, and std_k denotes the standard deviation of the k th band. We stack mea_k and std_k band by band, and the final spectral feature Spe can be expressed as follows: $\text{Spe} = \{\text{mea}_1, \text{std}_1, \dots, \text{mea}_k, \text{std}_k, \dots, \text{mea}_B, \text{std}_B\}$.

- 2) GLCM [43], [52] has been shown to efficiently describe the textural component of images. The gray tone of an image is generally 256 levels, which results in a large value of GLCM and a heavy computational load. Therefore, the gray level of the image is compressed to 8, and four Haralick's feature statistics [50] are used to describe the GLCM of each visual word in a compact way.

- a) Correlation

$$\text{Correlation} = \sum_{i,j=1}^L P_{ij} \frac{(i - \mu)(j - \mu)}{\sigma^2} \quad (5)$$

- b) Energy

$$\text{ASM} = \sum_{i,j=1}^L p_{ij}^2 \quad (6)$$

- c) Contrast

$$\text{Contrast} = \sum_{i,j=1}^L (i - j)^2 P_{ij} \quad (7)$$

- d) Homogeneity

$$\text{Homogeneity} = \sum_{i,j=1}^L \frac{P_{ij}}{1 + (i - j)^2} \quad (8)$$

In the aforementioned equations, μ and σ^2 denote the mean and variance of the GLCM, respectively, L denotes the image gray level, and P_{ij} denotes the i th line and j th column element of the normalized GLCM. The ultimate texture feature notation, in which cor_k , ene_k , con_k , and

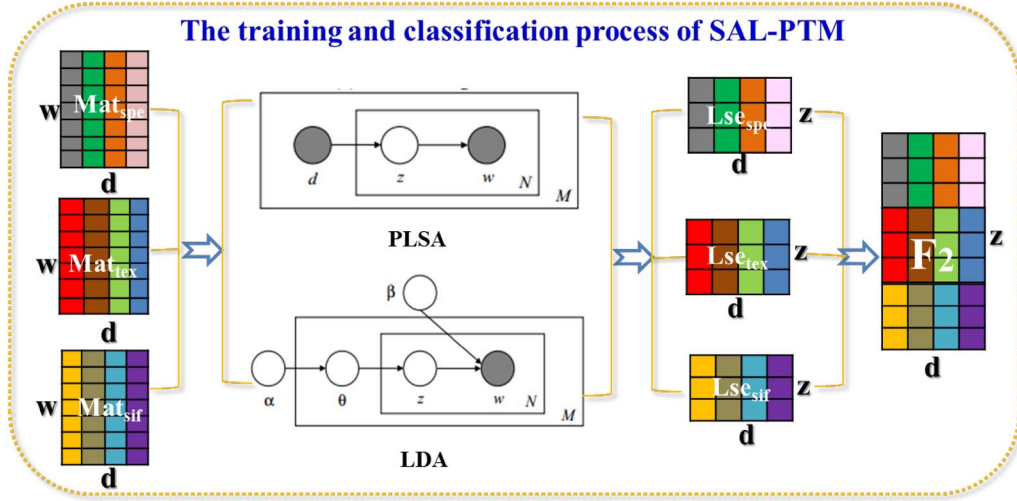


Fig. 5. Procedure of SAL-PTM-based HSR image scene classification.

hom_k represent the correlation, energy, contrast, and homogeneity of the k th band, respectively, can be represented as follows: $Tex = \{cor_1, ene_1, con_1, hom_1, \dots, cor_k, ene_k, con_k, hom_k, \dots, cor_B, ene_B, con_B, hom_B\}$.

- 3) Each visual word is segmented into 4×4 neighborhood sample regions for the SIFT feature. Eight directions for each gradient orientation histogram are counted in each sample region. Lowe [44] concluded that adopting a $4 \times 4 \times 8 = 128$ dimension vector to represent the keypoint descriptor could achieve an optimal effect. Hence, the final feature vector can then be represented by stacking the 128-dimension vectors $sif_k = \{sif_{k1}, \dots, sif_{k128}\}$ band by band as follows: $Sif = \{sif_1, \dots, sif_k, \dots, sif_B\}$.

In general, the strategy used in CAT-PTM is to simply concatenate Spe , Tex , and Sif . The image description is then denoted as $F_1 = \{Spe, Tex, Sif\}$, which is inadequate for determining HSR image scenes.

B. SAL Multifeature Fusion Strategy Based on PTM for HSR Image Scene Classification (SAL-pLSA and SAL-LDA)

On obtaining Spe , Tex , and Sif , they are quantized separately by k -means clustering, and three word-image co-occurrence matrices, Mat_{spe} , Mat_{tex} , and Mat_{sif} , are generated. SAL-PTM introduces probability statistics theory so that each element of the co-occurrence matrices is transformed into the word occurrence probability. For SAL-pLSA, it mines the latent semantic allocations of Mat_{spe} , Mat_{tex} , and Mat_{sif} according to (2). As a result, the latent semantic allocations Lse_{spe} , Lse_{tex} , and Lse_{sif} of Mat_{spe} , Mat_{tex} , and Mat_{sif} can be denoted as the mixed weights $p(z_k|d_i)$, where d_i represents a column in Mat_{spe} , Mat_{tex} , and Mat_{sif} . In this way, each of Lse_{spe} , Lse_{tex} , and Lse_{sif} can be represented as a set of vectors $\{p(z_1|d_i), \dots, p(z_K|d_i), \dots, p(z_K|d_i), d_i \in D\}$, where K denotes the topics selected for each of Mat_{spe} , Mat_{tex} , and Mat_{sif} . The final latent semantic allocation vector F_2 is acquired by concatenating Lse_{spe} , Lse_{tex} , and Lse_{sif} for all the images. SAL-LDA chooses a K -dimensional latent variable θ , and the probability distribution of each word w_n can be represented by a set of N topics z as $p(w_n|z_n, \beta)$, where β is a

$K \times V$ matrix following $\beta_{ij} = p(w^j = 1|z^i = 1)$ [26]. Given the parameters α and β , the joint distribution of a topic mixture θ , a set of N topics z , and a set of N words w is given by

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^N \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \quad (9)$$

where w indicates a column in Mat_{spe} , Mat_{tex} , and Mat_{sif} . The key problem when utilizing LDA is to compute the posterior distribution of the latent variables written by

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \quad (10)$$

However, due to the intractable computation, a variational inference algorithm with the Dirichlet parameter γ and the multinomial parameter (ϕ_1, \dots, ϕ_n) is employed to solve the problem in LDA. In more detail, suppose that there are N images; then, for each of Mat_{spe} , Mat_{tex} , and Mat_{sif} , K_{spe} , K_{tex} , and K_{sif} topics are selected, respectively, and the latent semantic allocations of Mat_{spe} , Mat_{tex} , and Mat_{sif} , denoted as Lse_{spe} , Lse_{tex} , and Lse_{sif} , respectively, are approximated by γ , as written by

$$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni}. \quad (11)$$

In the next procedure, Lse_{spe} , Lse_{tex} , and Lse_{sif} for all the images are adaptively concatenated according to each image, thus obtaining the final multifeature latent semantic allocation vector $F_2 = \{Lse_{spe}^T, Lse_{tex}^T, Lse_{sif}^T\}^T$, with the size of $(K_{spe} + K_{tex} + K_{sif}) \times N$ dimensions. Finally, the F_2 with the greatest discriminative capacity is input into the SVM classifier to predict the scene label of each class. The procedure of SAL-PTM is shown in Fig. 5.

IV. EXPERIMENTS AND ANALYSIS

In order to evaluate the different feature strategies and the different topic models (pLSA and LDA) for HSR image scene classification, a four-class USGS data set and the commonly used 21-class UC Merced data set were used in the scene

TABLE I
OPTIMAL K AND V VALUES FOR THE DIFFERENT FEATURE STRATEGIES AND DIFFERENT TOPIC MODELS

		SPECTRAL	TEXTURE	SIFT	CAT	SAL
pLSA	V	300	100	250	250	850
	K	19	17	15	17	57
LDA	V	300	300	300	150	750
	K	21	21	17	17	55

TABLE II
OVERALL CLASSIFICATION ACCURACIES FOR THE USGS DATA SET WITH pLSA AND LDA FOR THE DIFFERENT FEATURE STRATEGIES

	SPECTRAL	TEXTURE	SIFT	CAT	SAL
pLSA	94.60%	80.32%	83.81%	94.60%	96.19%
LDA	95.24%	83.81%	85.71%	94.89%	97.46%

TABLE III
OPTIMAL K AND V VALUES FOR THE DIFFERENT FEATURE STRATEGIES AND THE DIFFERENT TOPIC MODELS

		SPECTRAL	TEXTURE	SIFT	CAT	SAL
pLSA	V	800	800	1000	1000	3000
	K	70	70	70	70	205
LDA	V	2000	1000	1000	1000	3000
	K	70	80	70	100	210

TABLE IV
OVERALL CLASSIFICATION ACCURACIES FOR THE UC MERCED DATA SET WITH pLSA AND LDA FOR THE DIFFERENT FEATURE STRATEGIES

	SPECTRAL	TEXTURE	SIFT	CAT	SAL
pLSA	76.43%	78.33%	73.10%	76.79%	87.62%
LDA	76.67%	75.24%	72.86%	77.29%	88.33%

classification experiments. An original large image from the four-class USGS data set was also utilized to test the performance of SAL-PTM in an image annotation application. The single-feature methods and the conventional CAT-PTM were used for comparison. In addition, to further evaluate the combination and fusion strategies of SAL-PTM, the experimental results with the 21-class UC Merced data set, as published in the latest papers by Yang and Newsam in 2010 [53], Cheriyaad in 2014 [14], and Faria *et al.* in 2013 [38], are shown for comparison.

A. Experimental Setup for PTM-Based HSR Scene Classification

In this experiment, each image was empirically split into several 9×9 image patches. The 9×9 patches can be described by different features to digitize the image. To retain the spatial information between the adjacent image patches when conducting the even grid sampling, an overlap of three pixels was added, and thus, a better classification performance [34] was achieved. SVM with the RBF kernel was adopted to predict the scene labels of all the images. A k -means clustering 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 for

PTM. The visual word number V and topic number K were the two free parameters in our approach. Taking the computational complexity and the classification accuracy into consideration, V and K were optimally set as in Tables I and III for the different feature strategies and different topic models with the two data sets. In Tables I–IV, **SPECTRAL**, **TEXTURE**, and **SIFT** denote scene classification employing the mean and standard deviation-based spectral, GLCM-based texture, and SIFT-based structural features, respectively. **CAT** denotes the conventional multifeature fusion method which concatenates the spectral, texture, and SIFT features before k -means clustering. The proposed method that fuses the three features at the semantic latent allocation level is referred to as the **SAL** strategy. SVM was performed employing the LIBSVM package [54]. For the free parameters C and γ of SVM with the RBF kernel, a grid-search strategy was employed for the model selection. The multifeature latent semantic allocation vectors of the training data and testing data were put together and normalized between 0 and 1.

B. Experiment 1: Four-Class USGS Image Data Set

This experimental data set consists of 100 color aerial orthophotographs from the USGS, covering Montgomery County, Ohio, USA. These images mainly contain four scene classes:

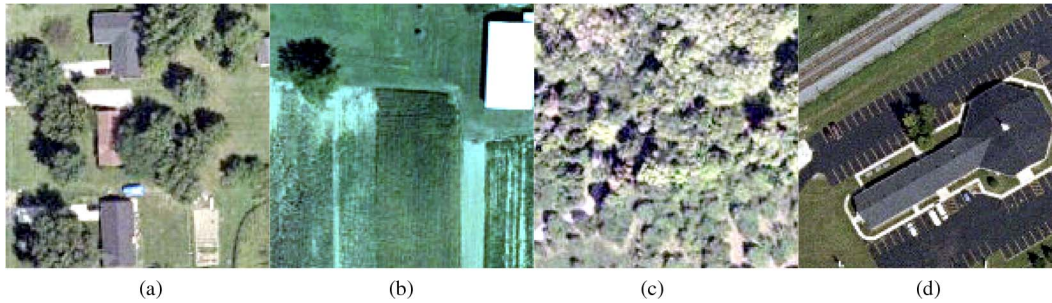


Fig. 6. Example images from the four-class USGS image data set. (a) Residential area. (b) Farm. (c) Forest. (d) Parking lot.

residential area, farm, forest, and parking lot, as shown in Fig. 6, with a spatial resolution of 2 ft. The original large images were split into a series of small experimental images with the size of 150×150 pixels. The four scene classes of residential area, farm, forest, and parking lot comprised 143, 133, 100, and 139 small images, respectively. A total of 50 images were randomly chosen from each scene class as the training samples, and the rest were used for testing.

1) *Evaluation of the Multiple Feature Extraction of SAL-PTM*: In Table II, the results of **SPECTRAL** for the pLSA model and LDA model are better than those of **TEXTURE** and **SIFT**. This is in accordance with the knowledge that the spectral feature is the fundamental information in HSR remote sensing images. However, the weakness is also apparent since the mean and standard deviation-based spectral feature cannot represent the complex spatial relationships between pixels in the patch. Hence, we chose GLCM [43] for the texture feature and SIFT [44] for the structural feature to combine with the spectral feature. It can be seen that the results obtained by **SAL** outperform **SPECTRAL**. This infers that the texture and SIFT features can compensate for the spectral feature, and the combination of the spectral, texture, and SIFT features can improve the scene classification performance of HSR images.

2) *Evaluation of the Proposed Procedure of SAL-PTM*: Table II shows the overall classification accuracies for the USGS data set with LDA and pLSA for the different feature strategies. The classification accuracies of **SAL** are 96.19% and 97.46% for the pLSA model and LDA model, respectively, which are higher than the classification accuracies of all the single-feature strategies and the conventional **CAT** at 94.60% and 94.89%. An interesting observation is that **SPECTRAL** performs no worse than **CAT**. This can be tolerated, as long as the result of **CAT** is better than the worst result of the single-feature strategies. This is mainly due to the inadequate fusion capacity of **k**-means clustering and the mutual interference between different features.

3) *Evaluation of the Different Topic Models Adopted by SAL-PTM*: We also compared the pLSA model with the LDA model. As can be seen from Table II, the classification results of LDA are slightly better than those of pLSA with the same feature strategy. LDA is a complete PTM, due to the definition of a probability function for the original discrete latent semantic allocation in pLSA. Compared with LDA, pLSA takes less time and resolves parameters more easily. This indicates that pLSA and LDA can both achieve good performances for HSR image scene classification.

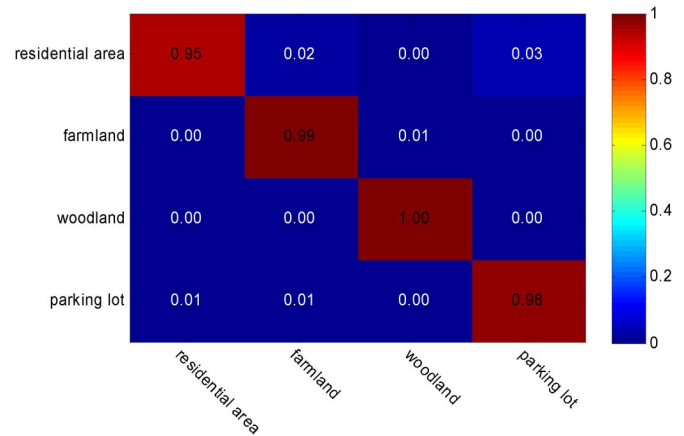


Fig. 7. Confusion matrix of SAL-LDA with the four-class USGS data set.

Fig. 7 displays the confusion matrix of SAL-LDA for the four-class USGS data set. As can be seen in the confusion matrix, there is some confusion between certain scenes. For instance, some scenes belonging to the residential area are classified as parking lot. This is, however, reasonable as there are many parking lots alongside the residential areas.

To allow a better visual inspection, some of the classification results of CAT-LDA and SAL-LDA are shown in Fig. 8.

C. Experiment 2: The UC Merced Image Data Set

To compare the scene classification performance of the proposed approach with the results of Yang and Newsam in 2010 [53] and Cheriadat in 2014 [14], we tested the classification performance with the challenging UC Merced data set. The challenging UC Merced evaluation data set was downloaded from the USGS National Map Urban Area Imagery collection [55]. It consists of 21 land-use scenes, which are manually labeled as follows: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts, as shown in Fig. 9. Each class separately consists of 100 images, which were cropped to 256×256 pixels, with a spatial resolution of 1 ft. The training samples were randomly selected from the UC Merced data set, and the remaining samples were retained for testing. The number of training samples per class was set from 20 to 80 with an interval of 10, to test the effect of the number of training samples per class, and the test results are given in Section V.



Fig. 8. Some of the classification results of CAT-LDA and SAL-LDA. The first, second, and third lines correspond to the scene classes of residential area, farm, and parking lot, respectively. (a) Correctly classified images for all the strategies. (b) Images which are classified correctly by SAL-LDA, but incorrectly classified by CAT-LDA.



Fig. 9. Example images from the 21-class UC Merced data set.

In this experiment, following the experimental setup in [55], 80 samples were randomly selected per class for training, and the remaining images were kept for testing. In addition, to further compare the proposed combination and fusion approach, the results of the feature combination and fusion method of Faria *et al.* [38] with the UC Merced data set were chosen for comparison. The color descriptors used in this experiment were border/interior pixel classification, the color coherence vector, and the global color histogram. The texture descriptors were the local activity spectrum and the quantized compound change histogram. The edge orientation autocorrelogram was used as

the shape descriptor. Six learning methods were also chosen: naive Bayes (NB), decision tree, NB tree, and three k -nearest neighbors (k -NN) strategies, using $k = 1$, $k = 3$, and $k = 5$. Faria *et al.* [38] performed the meta-learning methods with 36 classifiers (6 descriptors \times 6 learning methods), using the SVM classifier with the RBF kernel. The parameters in the experiment were set according to the recommendations of the authors. The experimental results are presented in Table V.

1) *Evaluation of the Multiple Feature Extraction of SAL-PTM*: As can be seen in Table IV, **SPECTRAL** does not always acquire the highest classification accuracies. The classification

TABLE V
COMPARISON WITH THE PREVIOUS REPORTED ACCURACIES FOR THE UC MERCED DATA SET

	SPECTRAL	TEXTURE	SIFT	CAT	SAL
pLSA	76.43%	78.33%	73.10%	76.79%	87.62%
LDA	76.67%	75.24%	72.86%	77.29%	88.33%

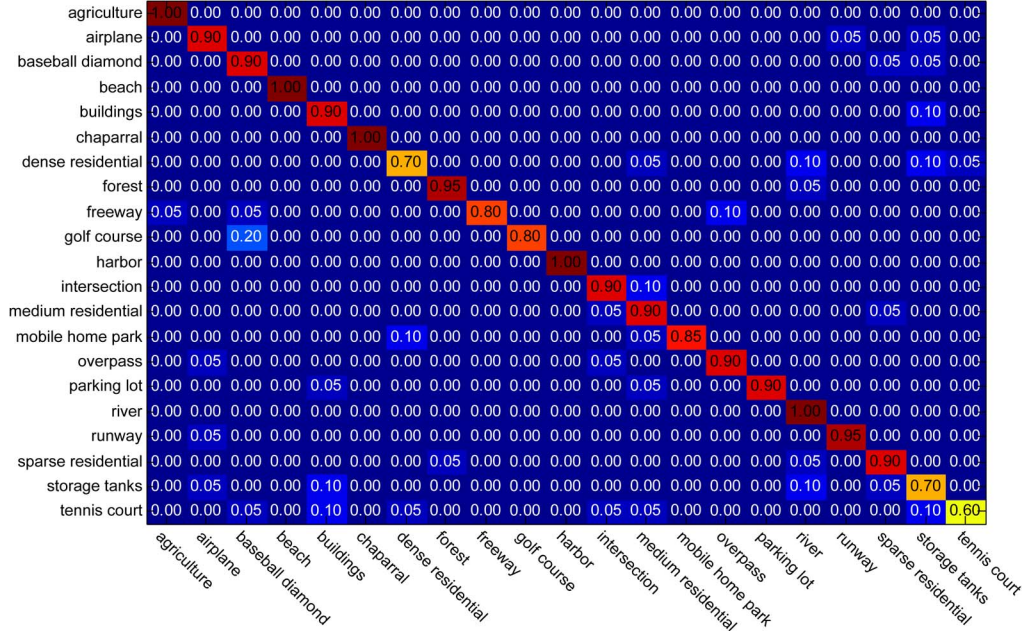


Fig. 10. Confusion matrix for SAL-LDA with the UC Merced data set.

result of **TEXTURE**, 78.33% for the pLSA model, is the best among the single-feature strategies. This is mainly due to the complex structure in the UC Merced data set. The results obtained by **SAL** are superior to that of the spectral strategies, which confirms the effectiveness of the combination of the spectral, texture, and SIFT features.

2) *Evaluation of the Proposed Procedure of SAL-PTM*: Among the five feature strategies, the classification performance of SAL, 87.62% and 88.33% for the pLSA model and LDA model, respectively, is the best, as shown in Table IV. These results are consistent with the analysis in *Experiment 1*. In Table V, it can be seen that, when compared to the best classification performances using the UC Merced data set, SAL-pLSA and SAL-LDA perform even better. When compared to the feature combination and fusion method of Faria *et al.* [38], SAL-pLSA and SAL-LDA can also obtain better results.

3) *Evaluation of the Different Topic Models Adopted by SAL-PTM*: As can be seen from Table IV, the PTM, including pLSA and LDA, can achieve good performances for HSR image scene classification. The classification results of pLSA for **TEXTURE** and **SIFT** are better than those of LDA. For the other strategies, LDA performs better. This indicates that neither LDA nor pLSA are superior to the other, and the classification performance depends on the experimental data set.

An overview of the performance of SAL-LDA is shown in the confusion matrix in Fig. 10. On the whole, most of the scene classes achieve good classification performances, and the agriculture, beach, chaparral, harbor, and river classes can be fully

recognized by SAL-LDA. There is some confusion between golf course and baseball diamond, dense residential and mobile home park, and freeway and overpass. This can be explained by the fact that the pairs of classes have similar spectral or structural features, such as both golf course and baseball diamond featuring vegetation cover and bare ground. In addition, it can be seen that some classes, such as storage tanks and airplane, have distinctive shape characteristics. Therefore, more work needs to be done with regard to the use of the shape feature.

To allow a better visual inspection, some of the classification results of CAT-LDA and SAL-LDA are shown in Fig. 11.

D. Experiment 3: Semantic Annotation of the Original Large Image

The size of the large image was $10\,000 \times 9\,000$ pixels, as shown in Fig. 12(a). In the annotation experiment, each class consisted of 50 training images with a size of 150×150 pixels. The large image was split into a set of small overlapping images of 150×150 pixels. In an empirical way, the overlap between two adjacent small images was set to 25 pixels. Hence, the spatial information lost during the large image sampling could be preserved. For the small images, the spectral, texture, and SIFT features also performed well when the patch size was set to 9×9 pixels and the overlap between two adjacent patches was set to three pixels. In the following procedure, all the small images were annotated with scene labels by the different feature strategies under the optimal parameter settings in *Experiment 1*.

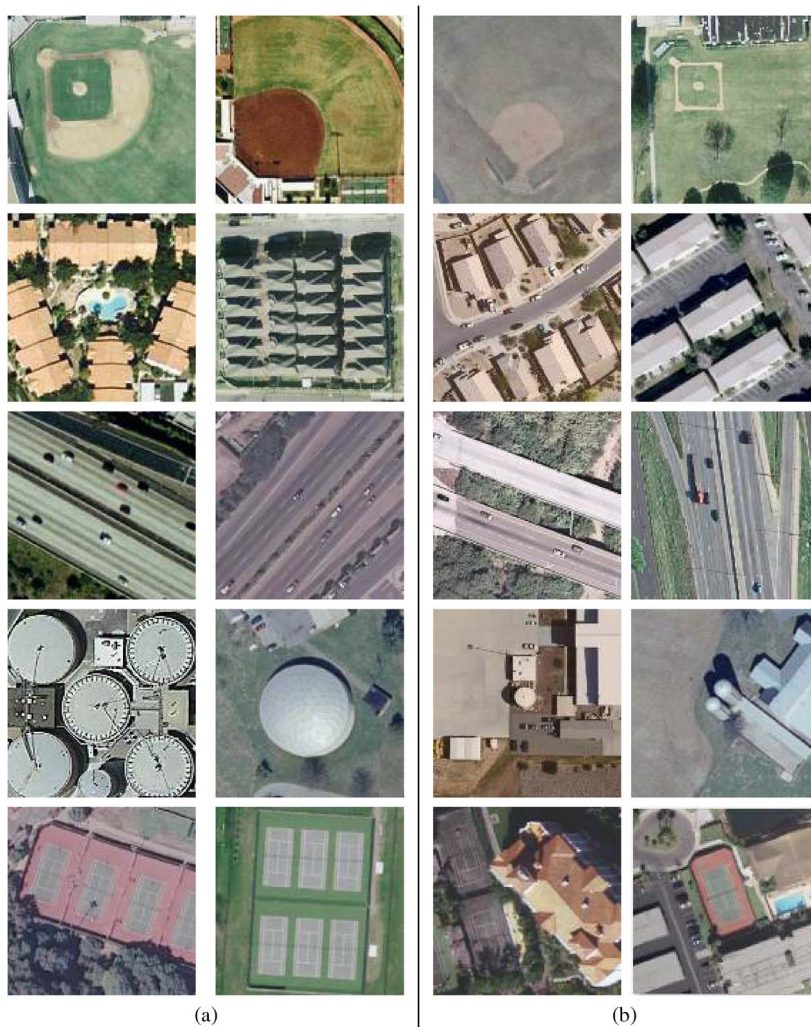


Fig. 11. Some of the classification results of CAT-LDA and SAL-LDA. The first, second, third, fourth, and fifth lines correspond to the scene classes of baseball diamond, dense residential, freeway, storage tanks, and tennis court, respectively. (a) Correctly classified images for all the strategies. (b) Images classified correctly by SAL-LDA, but incorrectly classified by CAT-LDA.

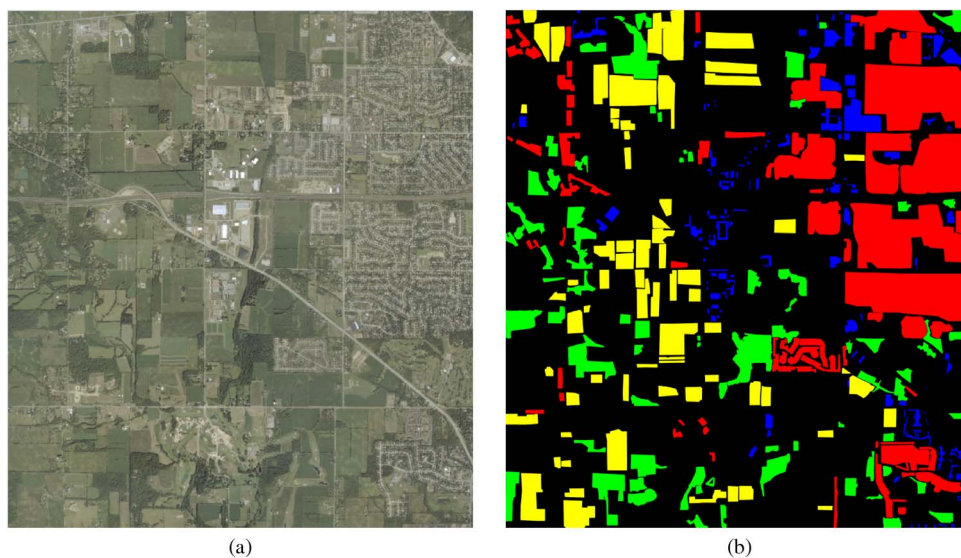


Fig. 12. Semantic annotation of the four-class USGS large image. (a) Original large image to be annotated. (b) Ground reference data.

TABLE VI
NUMBERS OF LABELED SAMPLES OF THE FOUR SCENES

Class name	Numbers of labeled samples
Residential area	13,539,173
Farm	7,009,696
Forest	6,141,878
Parking lot	2,189,729

To check whether the differences between the annotation results were meaningful, McNemar's test was employed to determine the statistical significance of the differences observed in two annotation results using the same test set [56]. Given two classifiers C_1 and C_2 (e.g., CAT-LDA and SAL-LDA), the number of pixels misclassified by C_1 but not by C_2 (M_{12}), and the number of cases misclassified by C_2 but not by C_1 (M_{21}) were compared. If $M_{12} + M_{21} \geq 20$, the X^2 [where X^2 is defined by (12)] statistic could be considered as following a chi-squared distribution with one degree of freedom [57]. Given a significance level of 0.05, $\chi_{0.05,1}^2 = 3.841459$, and if the McNemar's value was greater than $\chi_{0.05,1}^2$, the two classifiers were significantly different. The case of $M_{12} + M_{21} < 20$, for which the chi-squared approximation should not be applied, did not occur in the experiment

$$X^2 = \frac{(|M_{12} - M_{21}| - 1)^2}{M_{12} + M_{21}} \approx \chi_1^2. \quad (12)$$

A visual comparison and quantitative evaluation were considered to assess the semantic annotation results. To evaluate the annotation accuracy, a field map is provided in Fig. 12(b), based on ground reference data. Table VI provides the numbers of labeled samples of the four scenes. The spectral feature reflects the fundamental and important information in HSR images and was used as the annotation feature in [34]. Hence, the spectral feature strategy was selected as the single-feature strategy for the large image annotation. The final visual results of the scene annotation acquired by the SPECTRAL, CAT, and SAL strategies with the pLSA and LDA models (denoted as s-pLSA, CAT-pLSA, SAL-pLSA, s-LDA, CAT-LDA, and SAL-LDA) are shown in Fig. 13.

On the whole, most of the scene regions are annotated correctly, particularly the residential area scene and the farm scene. It can be seen that CAT-pLSA, CAT-LDA, SAL-pLSA, and SAL-LDA perform better than s-pLSA and s-LDA in the classification of the parking lots surrounded by the residential areas. Cut from Fig. 13(a)–(c), respectively, (g), (h), and (i) represent the same regions, to allow a detailed evaluation. In these three images, SAL-pLSA and CAT-pLSA can recognize the farm scene, while s-pLSA cannot. SAL-LDA improves the annotation performance in the distribution of residential area, which can be seen from the comparison of (j), (k), and (l). Among all the annotation results, SAL-pLSA and SAL-LDA obtain the best performances from the visual inspection. However, some misclassifications do occur, due to the unknown scene classes in the test images. For instance, the main roads are annotated as a parking lot scene or residential area scene. Hence, more work is needed to further define semantic concepts that cover all the possible area types or to add a reject class for the test image.

For the quantitative evaluation of the results, the corresponding annotation accuracies, evaluated by pixels, are presented in Table VII. From Table VII, it can be seen that the accuracies of SAL-LDA and SAL-pLSA for the residential scene and parking lot scene are higher than that of the other methods. CAT-pLSA and CAT-LDA obtain higher overall accuracies than s-pLSA and s-LDA, respectively, with SAL-pLSA and SAL-LDA obtaining the highest overall accuracies. This result infers that the proposed fusion procedures of **SAL** and the combination of the spectral, texture, and SIFT features in **SAL** both work effectively.

In addition to the annotation accuracy, Table VIII provides a pairwise comparison of the six methods using the McNemar's test for the USGS data set. In order to evaluate the statistical significance of the annotation results, we randomly selected about one-ninth of the test samples for the McNemar's test. As shown in Table VIII, all the McNemar's values of SAL-pLSA and SAL-LDA are greater than the critical value of $\chi_{0.05,1}^2$ (3.841459), which means that the differences are significant and SAL-PTM performs better than the other methods.

V. SENSITIVITY ANALYSIS

The number of visual words V generated by the k -means clustering plays a significant role in HSR image scene classification. Hence, a sensitivity analysis between the visual word number V and the four methods (CAT-pLSA, CAT-LDA, SAL-pLSA, and SAL-LDA) was carried out. In addition, to investigate the effectiveness of the PTM, as applied to scene classification, we analyzed the sensitivity of the topic number K mined by the PTM for the four methods. The effects of the number of training samples per class and the different classifiers for HSR image scene classification were also tested. The 21-class UC Merced data set was used for the sensitivity analysis.

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

To investigate the sensitivity of CAT-PTM (CAT-pLSA and CAT-LDA) and SAL-PTM (SAL-pLSA and SAL-LDA) in relation to parameter V , the values of the patch scale, the overlap, and the topic number K were kept constant at 9, 3, and 210, respectively. The visual word number V was then varied over the range of [1000, 1500, 2000, 2500, 3000, 3500] for the UC Merced data set.

As shown in Fig. 14, it can be clearly seen that SAL-PTM is superior to CAT-PTM over the entire range. With the increase in the visual word number V , the classification accuracy of CAT-PTM tends to decline, while the accuracy of SAL-PTM tends to increase. SAL-LDA obtains the highest accuracy of 88.33% when V is 3000. CAT-LDA provides an overall precision of 77.29% when V is 1000, with all of the methods being sensitive to the visual word number V .

B. Sensitivity Analysis in Relation to the Topic Number K

To study the sensitivity of CAT-PTM (CAT-pLSA and CAT-LDA) and SAL-PTM (SAL-pLSA and SAL-LDA) in relation

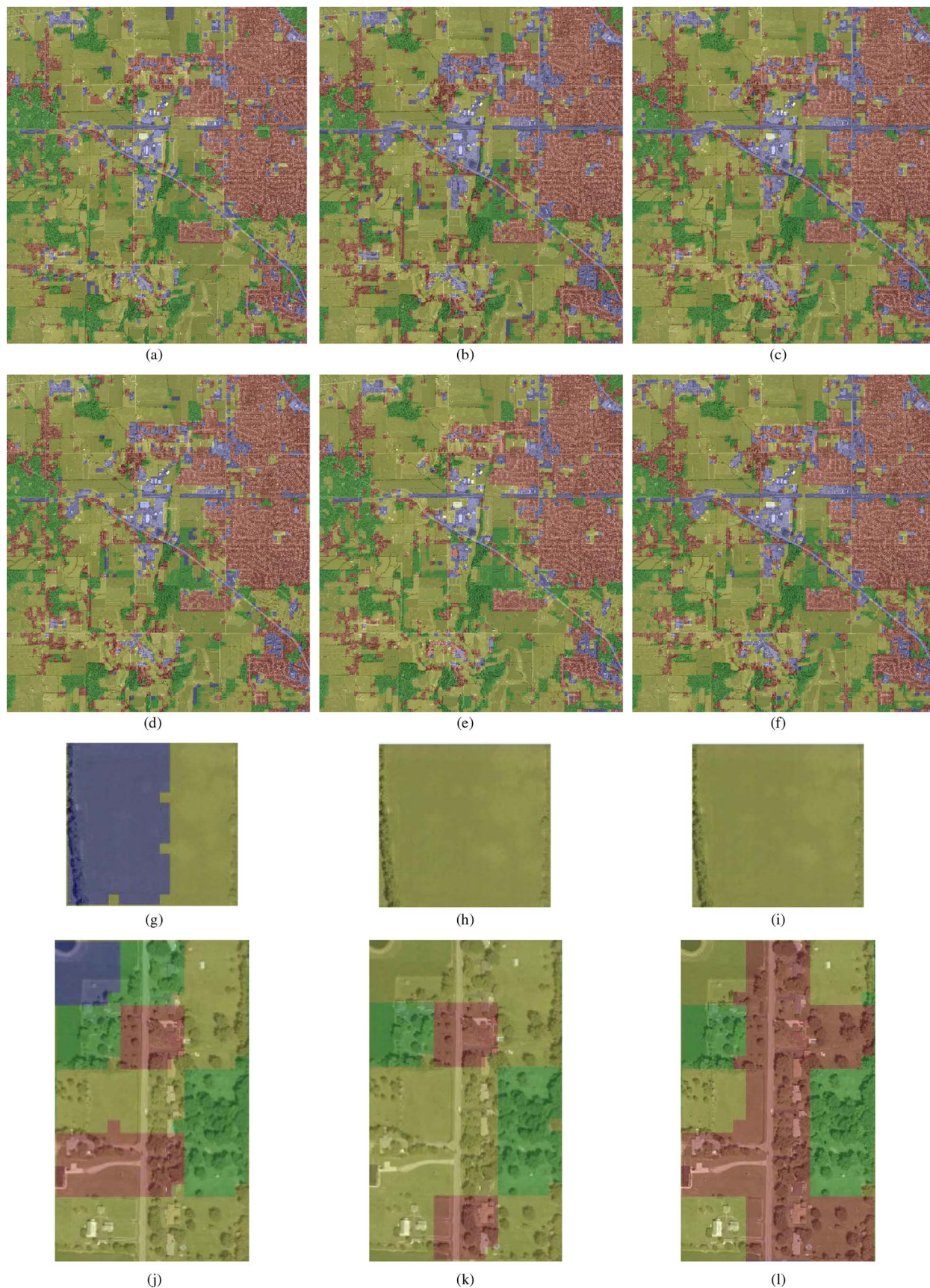


Fig. 13. Visual results of the semantic annotation of the original large image containing four semantic classes: (Brick red) Residential area, (yellow) farm, (green) forest, and (purple) parking lot, with the six methods (a) s-pLSA, (b) CAT-pLSA, (c) SAL-pLSA, (d) s-LDA, (e) CAT-LDA, and (f) SAL-LDA. (g), (h), and (i) are detailed comparisons of the same regions from (a), (b), and (c), respectively. (j), (k), and (l) are detailed comparisons of the same regions from (d), (e), and (f), respectively.

TABLE VII
ANNOTATION ACCURACIES WITH THE USGS DATA SET FOR THE DIFFERENT FEATURE STRATEGIES

Feature strategy	s-pLSA	CAT-pLSA	SAL-pLSA	s-LDA	CAT-LDA	SAL-LDA
Residential area	90.25%	92.12%	94.06%	92.09%	91.61%	95.79%
Farm	90.44%	91.28%	92.33%	92.59%	91.97%	91.60%
Forest	91.58%	82.04%	89.32%	85.67%	89.88%	87.82%
Parking lot	57.20%	79.28%	79.68%	68.27%	71.93%	77.47%
Overall accuracy	88.07%	88.80%	91.54%	89.04%	89.84%	91.69%

TABLE VIII
MCNEMAR'S TEST FOR THE USGS DATA SET

Methods	s-pLSA	CAT-pLSA	SAL-pLSA	s-LDA	CAT-LDA	SAL-LDA
s-pLSA	NA	5.94	192.68	3.34	9.73	195.10
CAT-pLSA		NA	153.10	4.70	2.79	160.17
SAL-pLSA			NA	188.04	159.01	5.07
s-LDA				NA	4.10	189.14
CAT-LDA					NA	147.00
SAL-LDA						NA

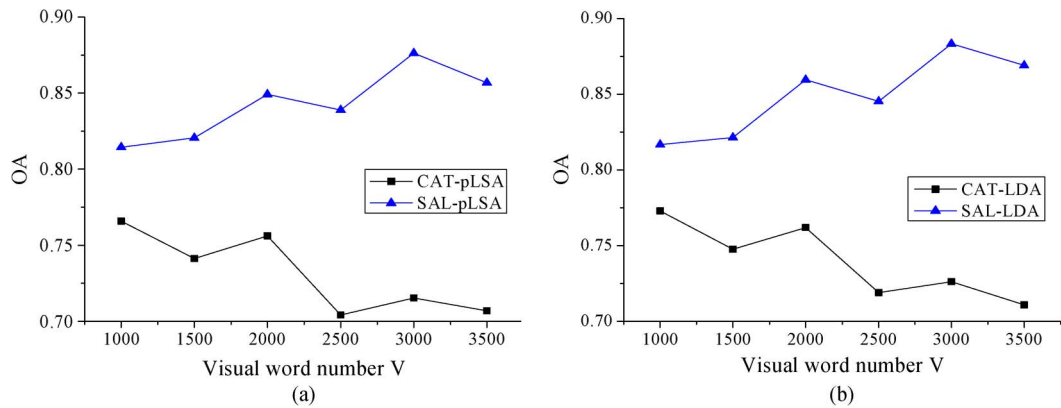


Fig. 14. Sensitivity analysis of CAT-PTM and SAL-PTM in relation to the visual word number V . (a) pLSA model. (b) LDA model.

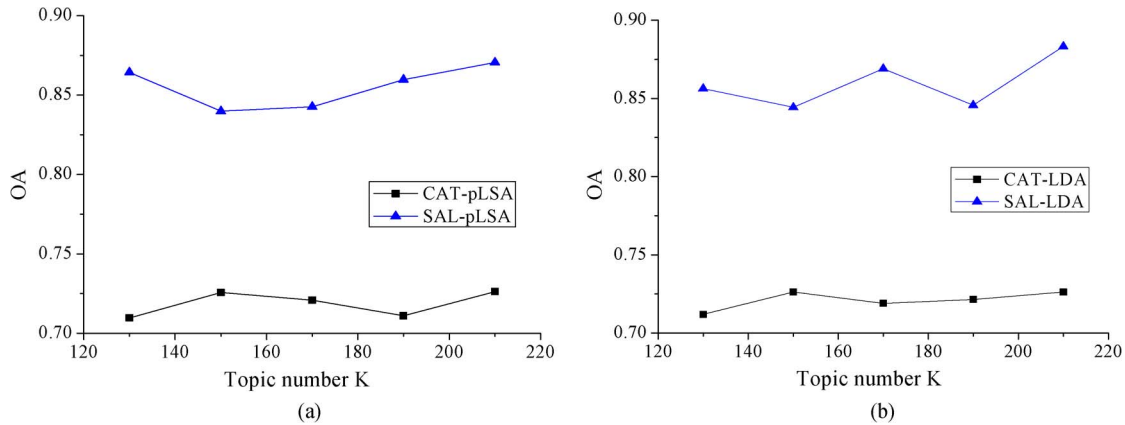


Fig. 15. Sensitivity analysis of CAT-PTM and SAL-PTM in relation to the topic number K . (a) pLSA model. (b) LDA model.

to parameter K , the values of the patch scale, the overlap, and the visual word number V were kept constant at 9, 3, and 3000. The topic number K was varied over the range of [120, 140, 160, 180, 200, 220] for the 21-class UC Merced data set.

From Fig. 15, it is notable that SAL-PTM obtains a remarkable performance, which far transcends that of the conventional CAT-PTM. Comparing Figs. 14 with 15, it can be clearly seen that the results of CAT-PTM and SAL-PTM display a greater fluctuation in relation to the visual word number V , and

CAT-PTM and SAL-PTM are less sensitive to the topic number K . Therefore, the number of topics can be set as fixed at first to determine the optimal number of visual words.

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

The effect of the number of training samples per class for CAT-PTM (CAT-pLSA and CAT-LDA) and SAL-PTM

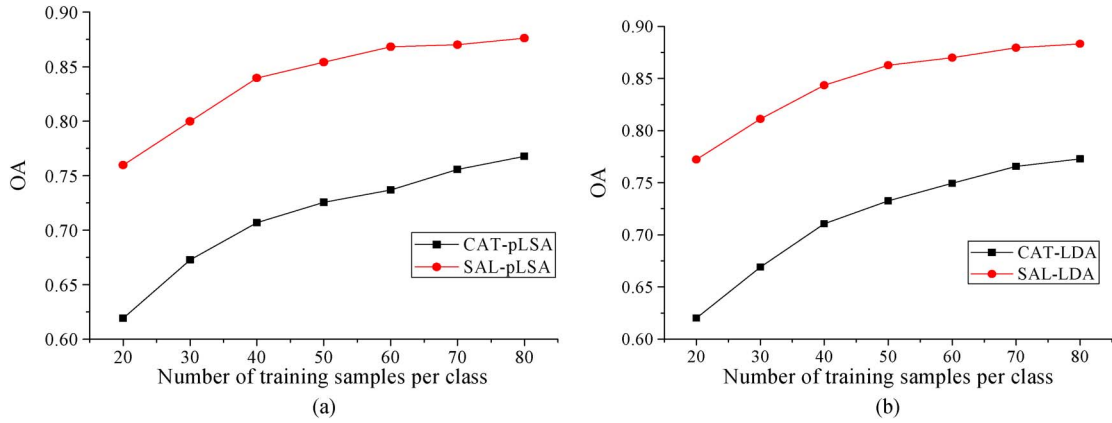


Fig. 16. Sensitivity analysis of CAT-PTM and SAL-PTM in relation to the number of training samples. (a) pLSA model. (b) LDA model.

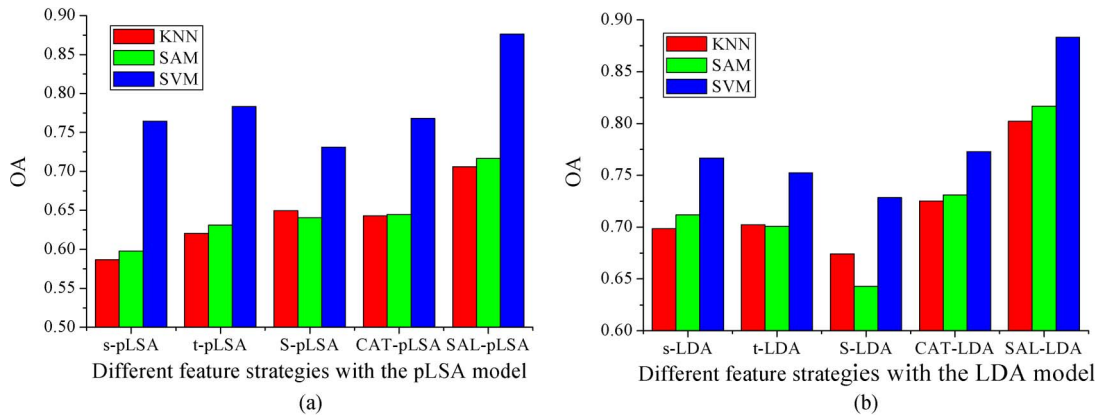


Fig. 17. Sensitivity analysis of the different methods in relation to different classifiers. (a) pLSA model. (b) LDA model.

(SAL-pLSA and SAL-LDA) was tested using the 21-class UC Merced data set. The number of training samples in each scene class was set from 20 to 80 with an interval of 10. In the experiment, the training samples were randomly selected from the data set, and the remaining images were kept for testing. As shown in Fig. 16, the proposed SAL-PTM performs better and is relatively stable with the increase in the number of training samples per class, when compared to CAT-PTM.

D. Sensitivity Analysis in Relation to Different Classifiers

In order to study the effect of different classifiers for the PTM-based HSR image scene classification, the **SPECTRAL**, **TEXTURE**, **SIFT**, **CAT**, and **SAL** strategies with the pLSA and LDA models (denoted as s-pLSA, t-pLSA, S-pLSA, CAT-pLSA, SAL-pLSA, s-LDA, t-LDA, S-LDA, CAT-LDA, and SAL-LDA) were tested using the k -NN, spectral angle mapper (SAM), and SVM classifiers, respectively. All the experimental results were obtained under the optimal parameter settings for the 21-class UC Merced data set. The values of K for k -NN were set to 3 for the **SPECTRAL** and **SAL** strategies, 1 for the **TEXTURE** and **SIFT** strategies, and 4 for the **CAT** strategy, respectively. It can be seen from Fig. 17 that SVM outperforms k -NN and SAM for all the methods, which confirms that SVM is an appropriate classifier for PTM-based HSR image scene classification.

VI. CONCLUSION

In this paper, an effective SAL multifeature fusion strategy based on PTM is proposed for HSR image scene classification (SAL-PTM). Considering the abundant information and complex structure in HSR images, SAL-PTM efficiently combines three complementary features. We chose the mean and standard deviation for the spectral feature, GLCM for the texture feature, and SIFT for the structural feature. The combination of the three features is able to capture the characteristics of HSR imagery.

In contrast to CAT-PTM, the three feature vectors in SAL-PTM are separately extracted and quantized by k -means clustering. This circumvents the inadequate fusion capacity of k -means clustering. The latent semantic allocations from the complete visual dictionary are mined by PTM separately and are then fused into the final semantic allocation vector. The proposed method is able to yield abundant low-level feature descriptions and adequate semantic allocations, and on this basis, accurate high-level concepts can be obtained for HSR imagery.

Experiments performed on a USGS data set and the UC Merced data set confirmed that the proposed SAL-PTM outperforms the conventional CAT-PTM and the single-feature methods. In our future work, we will consider other topic models which can relax the normalization constraint of the PTM. Moreover, other texture, shape, or structural features which are more appropriate for HSR images will be explored for HSR image scene classification.

ACKNOWLEDGMENT

The authors would like to thank the editor, associated editor, and anonymous reviewers for their helpful comments and Dr. F. A. Faria of the Federal University of Sao Paulo (UNIFESP), Brazil, for providing the experimental results of his algorithm. The authors would also like to thank Dr. J. D. Wegner and P. Tokarczyk of the Swiss Federal Institute of Technology Zürich, Switzerland, Prof. J. A. dos Santos of Universidade Federal de Minas Gerais, Brazil, and Prof. M. Wang of Nanjing Normal University, China, for their helpful suggestions to improve this paper.

REFERENCES

- [1] J. C. Tilton, Y. Tarabalka, P. M. Montesano, and E. Gofman, "Best merge region-growing segmentation with integrated nonadjacent region object aggregation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4454–4467, Nov. 2012.
- [2] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010.
- [3] I. A. Rizvi and B. K. Mohan, "Object-based image analysis of high-resolution satellite images using modified cloud basis function neural net-work and probabilistic relaxation labeling process," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4815–4820, Dec. 2011.
- [4] R. Bellens *et al.*, "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.
- [5] P. Gamba, F. Dell'Acqua, G. Lisini, and G. Trianni, "Improved VHR urban area mapping exploiting object boundaries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 8, pp. 2676–2682, Aug. 2007.
- [6] 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.
- [7] A. Bosch, X. Munoz, and R. Marti, "Which is the best way to organize/classify images by content?" *Image Vis. Comput.*, vol. 25, no. 6, pp. 778–791, Jul. 2007.
- [8] D. Tao, L. Jin, Z. Yang, X. Li, and L. Xuelong, "Rank preserving sparse learning for kinect based scene classification," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1406–1417, Oct. 2013.
- [9] B. Fernando, E. Fromont, and T. Tuytelaars, "Mining mid-level features for image classification," *Int. J. Comput. Vis.*, vol. 108, no. 3, pp. 1–18, Feb. 2014.
- [10] J. Luo, M. Boutell, R. T. Gray, and C. Brown, "Image transform bootstrapping and its applications to semantic scene classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 563–570, Jun. 2005.
- [11] 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-Verlag, 2013, pp. 324–333.
- [12] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.
- [13] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, Apr. 2012.
- [14] A. M. Cheriyyadath, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.
- [15] 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.
- [16] D. L. Wang and X. Liu, "Scene analysis by integrating primitive segmentation and associative memory," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 32, no. 3, pp. 254–268, Jun. 2002.
- [17] K. Koperski, G. Marchisio, S. Aksoy, and C. Tusk, "VisiMine: Interactive mining in image databases," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2002, vol. 3, pp. 1810–1812.
- [18] L. Zhou, Z. Zhou, and D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognit.*, vol. 46, no. 1, pp. 424–433, Jan. 2013.
- [19] A. Bolovinou, I. Partikakis, and S. Perantonis, "Bag of spatio-visual words for context inference in scene classification," *Pattern Recognit.*, vol. 46, no. 3, pp. 1039–1053, Mar. 2013.
- [20] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 381–392, Apr. 2011.
- [21] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1575–1589, Sep. 2007.
- [22] Y. Huang, K. Huang, C. Wang, and T. Tan, "Exploring relations of visual codes for image classification," in *Proc. IEEE CVPR*, 2011, pp. 1649–1656.
- [23] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2006, pp. 977–984.
- [24] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1/2, pp. 177–196, Jan. 2001.
- [25] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
- [26] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *Proc. Eur. Conf. Comput. Vis.*, May 2006, vol. 4, pp. 517–300.
- [27] B. Jin, W. Hu, and H. Wang, "Image classification based on pLSA fusing spatial relationships between topics," *IEEE Signal Process. Lett.*, vol. 19, no. 3, pp. 151–154, Mar. 2012.
- [28] 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.
- [29] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 524–531.
- [30] P. Quelhas *et al.*, "Modeling scenes with local descriptors and latent aspects," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, vol. 1, pp. 883–890.
- [31] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, vol. 1, pp. 370–377.
- [32] S. Chen and Y. L. Tian, "Evaluating effectiveness of latent Dirichlet allocation model for scene classification," in *Proc. 20th Annu. IEEE WOCC*, 2011, pp. 1–6.
- [33] C. Vaduva, I. Gavati, 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.
- [34] M. Liénou, H. Maître, 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.
- [35] K. Xu, W. Yang, G. Liu, and H. Sun, "Unsupervised satellite image classification using Markov field topic model," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 173–176, Jan. 2013.
- [36] M. Wang, Q. M. Wan, L. B. Gu, and T. Y. Song, "Remote-sensing image retrieval by combining image visual and semantic features," *Int. J. Remote Sens.*, vol. 34, no. 12, pp. 4200–4223, Jun. 2013.
- [37] 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.
- [38] F. A. Faria, J. A. dos Santos, A. R. Rocha, and R. da S. Torres, "A framework for selection and fusion of pattern classifiers in multimedia recognition," *Pattern Recognit. Lett.*, vol. 39, pp. 52–64, Apr. 2014.
- [39] J. A. dos Santos, P.-H. Gosselin, S. Phillip-Foliguet, R. da S. Torres, and A. X. Falcão, "Multiscale classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3764–3775, Oct. 2012.
- [40] 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.
- [41] W. Luo, H. Li, and G. Liu, "Automatic annotation of multispectral satellite images using author-topic model," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 4, pp. 634–638, Jul. 2012.
- [42] W. Luo, H. Li, G. Liu, and L. Zeng, "Semantic annotation of satellite images using author—Genre—Topic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1356–1368, Feb. 2014.
- [43] A. Baraldi and F. Parmiggiani, "An investigation of the textural characteristic associated with gray level cooccurrence matrix statistical parameters," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 2, pp. 293–304, Mar. 1995.
- [44] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [45] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

- [46] J. Suykens and J. Vanderwalle, "Least squares support vector machines classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
- [47] Z. Wu, Y. Huang, L. Wang, and T. Ta, "Group encoding of local features in image classification," in *Proc. IEEE Int. Conf. Pattern Recog.*, 2012, pp. 1505–1508.
- [48] J. Wang *et al.*, "Locality constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3360–3367.
- [49] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci. Inf. Eng., Nat. Taiwan Univ., Taipei, Taiwan, 2003.
- [50] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
- [51] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.
- [52] V. Sebastian, A. Unnikrishnan, and K. Balakrishnan, "Grey level co-occurrence matrices: Generalization and some new features," *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 2, no. 2, pp. 151–157, Apr. 2012.
- [53] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1465–1472.
- [54] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27, Apr. 2011.
- [55] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL GIS*, 2010, pp. 270–279.
- [56] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947.
- [57] O. Debeir, I. Van den Steen, P. Latinne, P. Van Ham, and E. Wolff, "Textural and contextual land-cover classification using single and multiple classifier systems," *Photogramm. Eng. Remote Sens.*, vol. 68, no. 6, pp. 597–606, Jun. 2002.



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 with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, since 2007 and is currently a Professor. His research interests include multi- and hyperspectral remote sensing data processing, high resolution image processing and scene analysis, and computational intelligence. He has published more than 70 research papers, including more than 30 peer-reviewed articles in international journals such as the *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, the *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART B*, and *Pattern Recognition*.

Dr. Zhong was the recipient of the National Excellent Doctoral Dissertation Award of China (2009) and New Century Excellent Talents in University of China (2009). He was a Referee of more than 20 journals, including the *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING* AND *PATTERN RECOGNITION*. He also serves as a Guest Editor of *Soft Computing*.



Qiqi Zhu (S'14) received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2013. She is currently working toward the Ph.D. degree in photogrammetry and remote sensing in the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

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



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 Chinese Academy of Sciences, Xian, 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. He is also a Chang-Jiang Scholar

Chair Professor appointed by the Ministry of Education of China. He is currently a Principal Scientist for the China State Key Basic Research Project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He has more than 360 research papers. He is the holder of 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, an Executive Member (Board of Governor) of the China National Committee of the International Geosphere-Biosphere Programme, and an Executive Member of the China Society of Image and Graphics. He regularly serves as a Cochair 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 Geo-spatial Information Science*, the *Journal of Remote Sensing*, and *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*.