SCENE CLASSFICATION BASED ON THE SEMANTIC-FEATURE FUSION FULLY SPARSE TOPIC MODEL FOR HIGH SPATIAL RESOLUTION REMOTE SENSING IMAGERY

Qiqi Zhu[a,b],Yanfei Zhong[a,b,*] , Liangpei Zhang[a,b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China – wuxi5477@126.com, (zhongyanfei, zlp62)@whu.edu.cn
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

Commission VII, WG VII/4

KEY WORDS: Scene classification, Fully sparse topic model, Semantic-feature, Fusion, Limited training samples, High spatial resolution

ABSTRACT:

Topic modeling has been an increasingly mature method to bridge the semantic gap between the low-level features and high-level semantic information. However, with more and more high spatial resolution (HSR) images to deal with, conventional probabilistic topic model (PTM) usually presents the images with a dense semantic representation. This consumes more time and requires more storage space. In addition, due to the complex spectral and spatial information, a combination of multiple complementary features is proved to be an effective strategy to improve the performance for HSR image scene classification. But it should be noticed that how the distinct features are fused to fully describe the challenging HSR images, which is a critical factor for scene classification. In this paper, a semantic-feature fusion fully sparse topic model (SFF-FSTM) is proposed for HSR imagery scene classification. In SFF-FSTM, three heterogeneous features—the mean and standard deviation based spectral feature, wavelet based texture feature, and dense scale-invariant feature transform (SIFT) based structural feature are effectively fused at the latent semantic level. The combination of multiple semantic-feature fusion strategy and sparse based FSTM is able to provide adequate feature representations, and can achieve comparable performance with limited training samples. Experimental results on the UC Merced dataset and Google dataset of SIRI-WHU demonstrate that the proposed method can improve the performance of scene classification compared with other scene classification methods for HSR imagery.

1. INTRODUCTION

The rapid development of earth observation and remote sensing techniques has led to large amount of high spatial resolution (HSR) images with abundant spatial and structural information. Some of the most popular approaches are the object-based and contextual-based methods which can achieve precise object recognition (Bellens et al., 2008; Rizvi and Mohan, 2011; Tilton et al., 2012). Nevertheless, the HSR scenes often contain diverse land-cover objects, such as road, lawn, and building. The same type of objects may vary in spectral or structural based low-level features. The different distribution of the same land-cover objects may obtain different type of semantic scenes. And the same type of scenes may consist of different types of simple objects. These methods which are based on the low-level features are unable to capture the complex semantic concepts of different scene images. This leads to the divergence between the low-level data and the high-level semantic information, namely the “semantic gap” (Bratasanu et al., 2011). It’s a big challenge to bridge the semantic gap for HSR imagery. Scene classification, which can automatically label an image from a set of semantic categories (Bosch et al., 2007), as an effective method has been receiving more and more attention (Yang and Newsam, 2010; Cheriyadat, 2014; Zhao et al., 2013; Zhao et al., 2016b; Zhao et al., 2016c). Among the various scene classification methods, the bag-of-visual-words (BOVW) model has been successfully applied to capture the high-level semantics of HSR scenes without the recognition of objects in object-based scene classification methods (Zhao et al,. 2014). Based on the BOVW model, the probabilistic topic model (PTM) represents the scenes as a random mixture of visual words. The commonly used PTM, such as probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) and latent Dirichlet allocation (LDA) (Blei et al., 2003) mine the latent topics from the scenes and have been employed to solve the challenges of HSR image scene classification (Bosch et al., 2008; Liénou et al., 2010; Văduva et al., 2013).

To acquire latent semantics, the feature descriptors captured from HSR images are critical for PTM. In general, a single feature is employed and is inadequate (Zhong et al., 2015). Multi-feature based scene classification methods have also been proposed (Shao et al., 2013; Zheng et al., 2013; Tokarczyk et al., 2015). Considering the distinct characteristics of the HSR images, the features should be carefully designed to capture the abundant spectral and complex structural information. In addition, the different features are usually fused before *k*-means clustering, thus acquiring one dictionary and one topic space for all the features. This leads to the mutual interference between different features (Zhong et al., 2015), and is unable to circumvent the inadequate clustering capacity of the hard-assignment based *k*-means clustering, which is not efficient in the high-dimensional feature space. With the development of PTM for HSR scene classification, there are two issues should be considered. The first one is how to infer sparser latent representations of the HSR images. Another one is how to design more efficient inference algorithms for PTM. In order to achieve good performance for huge volume of HSR image

[*] Corresponding author

scenes, we may have to increase the number of topics to get more semantic information. However, for instance, the distribution of topic variables for the LDA model is drawn from a Dirichlet distribution with the parameter α . The variable is greater than 0 no matter how the parameter α varies (Blei et al., 2003). This leads to a dense topic representation of the HSR images, which is not sparse and requires more storage, and is time consuming. Another method is to impose sparsity constrains on the topic to change the object function of the model (Shashanka et al., 2007; Zhu and Xing, 2011). But we have to do model selection with the regularization terms based auxiliary parameters of these model, which is problematic when dealing with large amount of HSR image dataset. Fivefold cross validation is often performed to evaluate the experimental dataset to guarantee enough training samples for classification accuracy (Yang and Newsam, 2010; Cheriyadat, 2014). Reducing the number of training samples would be more practical.

Inspired by the aforementioned work, we present a semantic-feature fusion fully sparse topic model (SFF-FSTM) for HSR image scene classification. Fully sparse topic model (FSTM) proposed by Than and Ho (2012) for modeling large collections of documents is utilized to model HSR imagery for the following reason. Based on the similarity of documents and images, FSTM is able to remove the redundant information and infer sparse semantic representations with shorter inference time. In this way, to acquire sparse latent topics, we intended to use a limited number of images as training sample which is more in line with the practical application. To the best of our knowledge, no such PTM based scene classification method with limited training samples has been developed to date. However, FSTM is unable to fully exploit the information provided by the limited training samples with sparse representations. Hence in SFF-FSTM, three complementary features are selected to describe HSR images. Dense scale-invariant feature transform (SIFT) feature is chosen as the structural feature, mean and standard deviation as the spectral feature, and wavelet feature as the texture feature. Based on the effective feature description for HSR imagery, a semantic-feature fusion strategy is designed to fuse the three features after semantic mining with three distinct topic spaces. This can provide fully mined semantic information of the HSR imagery from three complementary perspectives, with no mutual interference and clustering impact. The incorporation of support vector machine (SVM) with a histogram intersection kernel (HIK) is effective in increasing the discrimination of different scenes. The combination of multiple semantic-feature fusion strategy and sparse representation based FSTM is able to trade off sparsity and the quality of sparse inferred semantic information as well as inferring time, and presents a comparable performance with the existed relevant method.

The rest of the paper is organized as follows. The next section details the procedure of the proposed SFF-FSTM for HSR image scene classification. A description of the experimental datasets and an analysis of the experimental results are presented in Section 3. Conclusions are discussed in the last section.

2. SEMANTIC-FEATURE FUSION FULLY SPARSE TOPIC MODEL FOR HSR IMAGERY

2.1 Probabilistic Topic model

Based on "bag-of-words" assumption, the generative probabilistic model of PTM, including PLSA, LDA and FSTM, are applied to HSR images by utilizing a visual analog of a word, acquired by vector quantizing spectral, texture, and structural feature like region descriptors (Bosch et al., 2008). Each image can then be represented as a set of visual words from the visual dictionary. By introducing the latent topics characterized by a distribution over words, the PTM model the images as random mixtures over latent variable space.

Among the various PTM, the PLSA model as the classical PTM is proposed by Hofmann (2001). It combines probability and statistics theory with the BOVW representation. By choose a latent topic z_k with probability $p(z_k \mid d_i)$ and a word w_j with probability $p(w_j|z_k)$, the joint probability $p(w_j, d_i)$ between visual words w_j and images d_i can be decomposed as (1):

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

The mixing weight $p(z_k \mid d_i)$ is the semantic information which PTM mined from the visual words of HSR images. It can be seen that PLSA lack a probability function to describe the images. This makes PLSA unable to assign probability to the images outside the training samples, and the number of model parameter grow linearly with the size of image dataset.

Hence, in 2003, Blei proposed LDA, which introduces the Dirichlet distribution over the topic mixture θ based on the PLSA model. The *k*-dimensional random variable θ follows the Dirichlet distribution with the parameter α , where k is assumed known and fixed first. The LDA model provides a probability function for the discrete latent topics in PLSA, which being a complete PTM. However, the Dirichlet variable is greater than 0 when α varies. The latent representation of HSR imagery by LDA is often dense with the large amount of images to model, while requiring huge memory for storage. And the inference algorithm of the LDA model is complex and takes a lot of time.

In 2012, Than and Ho proposed FSTM for modeling large collections of documents and applying to supervised dimension reduction. FSTM uses the Frank-Wolf algorithm of the sparse approximation algorithm as the inference algorithm, which follows the greedy approach, and has been proven to converge at a linear rate to the optimal solutions. In FSTM, the latent topic proportion θ is a convex combination of the topic simplex with at most *l*+1vertices after *l* iterations, which follows an implicit constraint $\|\theta\|_0 \leq L+1$. Hence, we choose FSTM with the sparse solutions to model the HSR imagery in this paper.

2.2 Complementary feature description

As can be seen from Fig. 1(a), it is difficult to distinguish parking lot from harbor, neither from the structural characteristics nor the textural ones. However, due to the spectral difference between ocean and road, the spectral characteristics play an important role. In Fig. 1(b), the storage tanks and dense residential scenes mainly differ in the structural characteristics. In addition, it can be seen from Fig. 1(c) that the forest and agriculture scenes are similar in spectral and structural characteristics, but they differ greatly in the textural information from the global perspective. Considering the abundant spectral characteristics and the complex spatial arrangement of HSR imagery, three complementary features are designed for the HSR imagery scene classification task. Before feature descriptor extraction, the images are split into image patches using uniform grid sampling method.

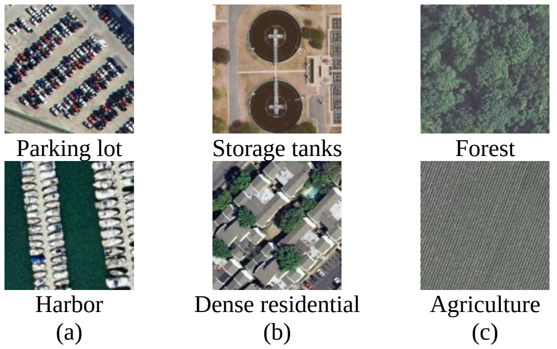


Fig. 1. HSR images of the parking lot, harbor, storage tanks, dense residential, forest, and agriculture scene classes: (a) shows the importance of the spectral characteristics for HSR images; (b) shows the importance of the structural characteristics for HSR images; and (c) shows the importance of the textural characteristics for HSR images.

2.2.1 Spectral feature: 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 (2) and (3), *n* is the total number of image pixels in the sampled patch, and v_{ij} denotes the *j*-th band value of the *i*-th 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.

$$mean_j = \frac{\sum_{i=1}^{n} v_i}{n} \tag{2}$$

$$std_j = \sqrt{\frac{\sum_{i=1}^{n} (v_{ij} - mean_j)^2}{n}} \tag{3}$$

2.2.2 Texture feature: The texture feature contains information about the spatial distribution of tonal variations within a band (Haralick et al., 1973), which can give consideration to both the macroscopic properties and fine structure. Wavelet transforms enable the decomposition of the image into different frequency sub-bands, similar to the way the human visual system operates (Huang and Avivente, 2008). This makes it especially suitable for image classification and multilevel 2-D wavelet decomposition is utilized to capture the texture feature from the HSR images. And the level where the wavelet decomposition of the images at is optimally set to 3.

2.2.3 Structural feature: The SIFT feature (Lowe, 2004) has been widely applied in image analysis since it can overcome the addition of noise, affine transformation, and changes in the illumination, as well as compensating for the deficiency of the spectral feature for HSR imagery. Each image patch is split into 4×4 neighbourhood regions and each directions for each gradient orientation histogram are 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 (Li and Perona, 2006), and Lowe (2004) suggest that using a $4 \times 4 \times 8 = 128$ dimensions vector to describe the keypoint descriptor is optimal.

2.3 Multiple Semantic-feature Fusion Fully Sparse Topic Model for HSR Imagery with Limited Training Samples

The previous studies have shown that a uniform grid sampling method can be more effective than other sampling methods such as random sampling (Li and Perona, 2006). In this way, the image patches acquired by uniformly sampling the HSR images are digitized by spectral, texture and SIFT features, and three types of feature descriptors, D_1, D_2, and D_3 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 values. The *k*-means clustering is applied to quantize the feature descriptors to generate 1-D frequency histogram, and image patches with similar feature values can correspond to the same visual word. By the statistical analysis of the frequency for each visual word, we can obtain the corresponding visual dictionary.

The conventional methods usually directly concatenate three types of feature descriptors to make up a long feature $F_1 = \{D_1, D_2, D_3\}$. The long vector is then quantized by *k*-mean clustering to generate a 1-D histogram for all the features. As the features interfere with each other when clustering, the 1-D histogram is unable to fully describe the HSR imagery. In SFF-FSTM, the spectral, texture, and SIFT features are quantized separately by k-mean clustering algorithm to acquire three distinct 1-D histograms, H_1, H_2, and H_3. By introducing probability theory, each element of the 1-D histogram for SFF-FSTM are transformed into the word occurrence probability. To mine the most discriminative semantic feature, which is also the core idea of PTM, the three histograms are separately mined by SFF-FSTM to generate three distinct latent topic spaces. This is different from the conventional strategies which fuse the three histograms before topic modeling, and only one latent topic space is obtained which is inadequate.

Specifically, SFF-FSTM chooses a *k*-dimensional latent variable θ . Given an image ***M*** and *K* topics $\beta = (\beta_1, ..., \beta_K)$, the log likelihood of ***M*** is defined in (4), where I_M is the set of term indices of image ***M***, and M_j is the frequency of term *j* in ***M***. Hence, the inference task is to search for θ to maximize the likelihood of ***M***. we can obtain (5) to set $x_j = \sum_{k=1}^{K} \theta_k \beta_{kj}$ and $x = (x_1, ..., x_V)^t$, where *V* is the visual dictionary of *V* terms. Different from other topic models, SFF-FSTM do not infer θ directly, whereas reformulate the inference task of optimization over θ as a concave maximization problem over the simplex $\Delta = conv(\beta_1, ..., \beta_K)$ of topic. It can be seen that ***x*** is a convex combination of the K topics with the fact in (6), and by finding $x \in \Delta$ that maximizes the objective function (5), we can infer the latent topic proportion of the image ***M***.

Hence, suppose there are N images, then for each of H_1, H_2, and H_3, K_1, K_2, and K_3 topics are assumed to compose the images, respectively. The latent semantics of H_1, H_2, and H_3, denoted as θ_1, θ_2, and θ_3, respectively , are inferred with the Frank-Wolf algorithm. Then the semantic features θ_1, θ_2, and θ_3 of all the HSR images are fused at the semantic level, thus obtaining the final multiple semantic-feature $F_2 = \{\theta_1^{\mathrm T}, \theta_2^{\mathrm T}, \theta_3^{\mathrm T}\}^{\mathrm T}$, with a sparse size. Finally, the F_2 with the optimal discriminative characteristics is classified by SVM classifiers with a HIK to predict the scene label. The HIK measures the degree of similarity between two histograms, to deal with the scale changes, and has been applied to image classification using color histogram features (Barla et al., 2003). We let $\tilde{\mathbf V} = (\tilde{\mathbf v}_1, \tilde{\mathbf v}_2, ..., \tilde{\mathbf v}_M)$ be the LGFBOVW representation vectors of M images, and the HIK is calculated according to (7). In this way, SFF-FSTM provides a complementary feature description, an effective image representation strategy, and an adequate topic modeling procedure for HSR image scene classification, with even limited training samples, which will be tested in the Experimental Section. Scene classification based on SFF-FSTM is shown in Fig.2.

$$\log P(M) = \sum\nolimits_{j\in I_M} M_j \log \sum\nolimits_{k=1}^{K} \theta_k \beta_{kj} \tag{4}$$

$$\log P(M) = \sum_{j\in I_M} M_j \log x_j \tag{5}$$

$$\sum\nolimits_k \theta_k = 1，\ \theta_k \ge 0 \tag{6}$$

$$K(\tilde{\mathbf v}_i, \tilde{\mathbf v}_j) = \sum_k \min(\tilde{\mathrm v}_{i,k}, \tilde{\mathrm v}_{j,k}) \tag{7}$$

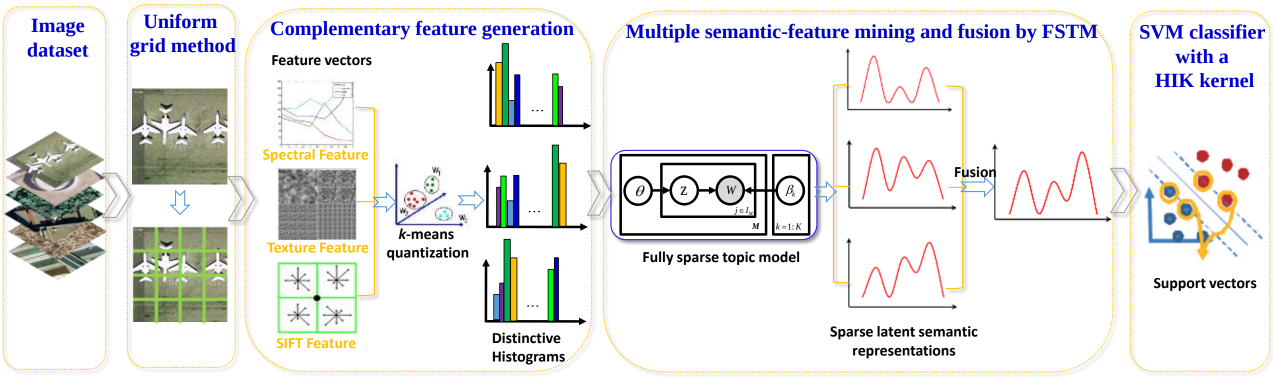


Figure 2. The proposed HSR scene classification based on the SFF-FSTM.

3. EXPERIMENTS AND ANALYSIS

3.1 Experimental Design

The commonly used 21-class UC Merced Dataset and a 12-class Google dataset of SIRI-WHU were evaluated to test the performance of SFF-FSTM. 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 LGFBOVW, the different methods were executed 100 times by a random selection of training samples, to obtain convincing results for the two datasets. 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 FSTM. The visual word number V and topic number K were the two free parameters in our method. Taking the computational complexity and the classification accuracy into consideration, V and K were optimally set as in Table 1 and Table 3 for the different feature strategies with the two dataset. In Table 1, 2, 3, and 4, SPECTRAL, TEXTURE, and SIFT denote scene classification utilizing the mean and standard deviation based spectral, wavelet-based texture, SIFT-based structural features, respectively. The proposed method that fuse the multiple semantic features at the latent topic level is referred to as the SFF strategy.

To further evaluate the performance of SFF-FSTM, the experimental results utilizing SPM (Lazebnik et al., 2006), PLSA (Bosch et al., 2008), LDA (Li énou et al., 2010) and the experimental results on the UC Merced dataset, as published in the latest papers by Yang and Newsam (2010), Cheriyadat (2014), Chen and Tian (2015), Mekhalfi *et al.* (2015), and Zhao *et al.* (2016a) are shown for comparison. SPM employed dense gray SIFT, and the spatial pyramid layer was optimally selected as one. In addition, the experimental results on the Google dataset of SIRI-WHU utilizing SPM (Lazebnik et al., 2006), PLSA (Bosch et al., 2008), LDA (Li énou et al., 2010) and the experimental results on the Google dataset of SIRI-WHU, as published in the latest paper by Zhao *et al.* (2016a) are also shown for comparison.

3.2 Experiment 1: The UC Merced Image Dataset

The UC Merced dataset was downloaded from the USGS National Map Urban Area Imagery collection (Yang and Newsam, 2010). This dataset consists of 21 land-use scenes (Fig. 3), 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, measuring 256×256 pixels, with a 1-ft spatial resolution. Following the experimental setup as published in Yang et al. (2010), 80 samples were randomly selected per class from the UC Merced dataset for training, and the rest were kept for testing.

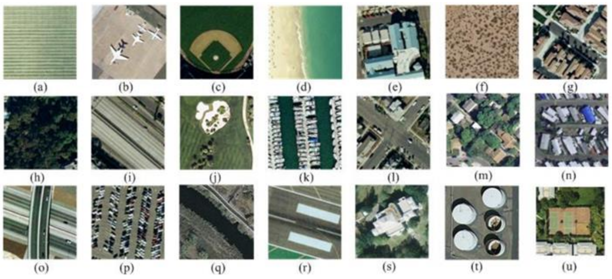


Figure 3. UC Merced dataset. (a)–(u): agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts.

The classification performance of different strategies based on the FSTM and the comparison with the experimental results of previous methods for the UC Merced dataset are reported in Table 2. As can be seen from Table 2, the classification results of the single feature based FSTM is unsatisfactory. The classification result, 94.55% ± 1.02% for the proposed SFF-FSTM is best among the different methods, and improves a lot compared with the single feature strategy. This indicates that the combination of multiple semantic-feature fusion strategy and sparse representation based FSTM is able to trade off sparsity and the quality of sparse inferred semantic information as well as inferring time. In addition, it can be seen that SFF-FSTM is superior to the performance of SPM (Lazebnik et al., 2006), PLSA (Bosch et al., 2008), LDA (Li énou et al., 2010), the Yang and Newsam method (2010), the Cheriyadat method (2014), the Chen and Tian method (2015), the Mekhalfi *et al.* method (2015), and the Zhao *et al.* method (2016a).

	SPECTRAL	TEXTURE	SIFT	SFF
V	1000	800	1000	2800
K	240	300	280	820

Table 1. Optimal *V* and *K* values for the different feature strategies with the UC Merced dataset

	SPECTRAL	TEXTURE	SIFT
FSTM	78.33±1.42	75.00±1.63	82.38 ±1.58
SPM	82.30±1.48		
PLSA	89.51±1.31		
LDA	81.92±1.12		
Cheriyadat (2014)	81.67±1.23		
Yang and Newsam (2010)	81.19		
Chen and Tian (2015)	89.10		
Mekhalfi *et al*. (2015)	94.33		
Zhao *et al*. (2016a)	92.92±1.23		
SFF-FSTM	94.55±1.02		

Table 2. Classification accuracies (%) of different strategies based on the FSTM and comparison with the experimental results of previous methods for the UC Merced dataset

3.3 Experiment 2: The Google Dataset of SIRI-WHU

The Google dataset was acquired from Google Earth (Google Inc.), covering urban areas in China, and the dataset is designed by Intelligent Data Extraction and Analysis of Remote Sensing (RS_IDEA) Group in Wuhan University (SIRI-WHU) (2016a). It consists of 12 land-use classes, which are labelled as follows: meadow, pond, harbor, industrial, park, river, residential, overpass, agriculture, water, commercial, and idle land, as shown in Fig. 4. 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 dataset, and the remaining samples were retained for testing.

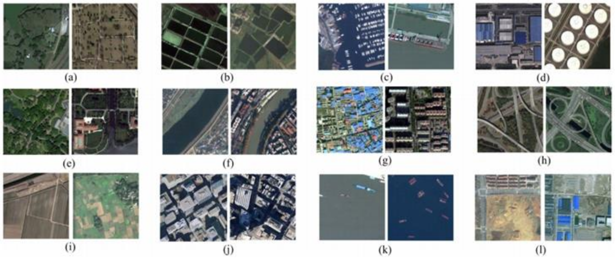


Figure. 4. Google dataset of SIRI-WHU. (a)–(l): meadow, pond, harbor, industrial, park, river, residential, overpass, agriculture, water, commercial, and idle land.

The classification performance of different strategies based on the FSTM and comparison with the experimental results of previous methods for the Google dataset of SIRI-WHU are reported in Table 4. As can be seen from Table 4, the classification results, 97.83%±0.93%, for the proposed SFF-FTSM, is much better than the spectral, texture, SIFT based FSTM method, which confirms the framework incorporating multiple semantic-feature fusion and FSTM is a comparative approach for HSR image scene classification. In Table 4, compared to the other methods, SPM, the LDA method proposed by Lienou *et al*. (2010), the PLSA method proposed by Bosch *et al*. (2008), and the experimental results published by Zhao *et al.* (2016a), the highest accuracy is required by the proposed SFF-FSTM, which presents a comparable performance with the existed relevant method.

	SPECTRAL	TEXTURE	SIFT	SFF
V	1000	800	1000	2800
K	240	300	280	820

Table 3. Optimal *V* and *K* values for the different feature strategies with the Google dataset of SIRI-WHU

	SPECTRAL	TEXTURE	SIFT
FSTM	83.33±1.06	80.92±0.95	78.50±1.12
SPM	77.69±1.01		
PLSA	89.60±0.89		
LDA	60.32±1.20		
Zhao et al. (2016a)	91.52±0.64		
SFF-FSTM	97.83±0.93		

Table 4. Classification accuracies (%) of different strategies based on the FSTM and comparison with the experimental results of previous methods for the Google dataset of SIRI-WHU

3.4 Experiment 3: Multiple Semantic-feature Fusion Fully Sparse Topic Model for HSR Imagery with Limited Training Samples

By modeling the large collection of images with only a few latent topic proportions of non-zero values, we intend to deal with the HSR imagery with limited training samples employing SFF-FSTM and SAL-LDA (Zhong et al., 2015), respectively. The training number was varied over the range of [80, 60, 40,

20, 10, 5] for the UC Merced dataset. And the training number for the Google dataset of SIRI-WHU was varied over the range of [100, 80, 60, 40, 20, 10]. The classification accuracy with different numbers of the training samples for the UC Merced dataset and the Google dataset of SIRI-WHU are reported in Table 5 and Table 6. The corresponding curves are shown in Fig. 5.
As can be seen from Table 5, Table 6 and Fig.5, the proposed SFF-FSTM performs better, and is relatively stable with the decrease in the number of training samples per class for the two datasets, when compared to SAL-LDA. When the training samples is under 20%, even 10% or 5%, SFF-FSTM display a smaller fluctuation than SAL-LDA, and can keep a comparative satisfactory and robust performance with limited training samples.

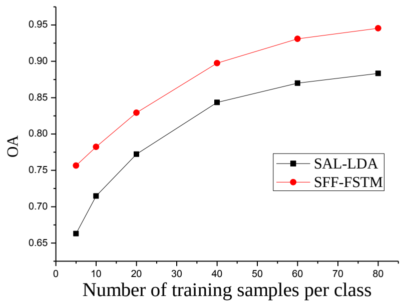
We also test and compare the inference efficiency of the proposed SFF-FSTM and SAL-LDA with the spectral feature for the two datasets. The inference time of SFF-FSTM is about 3 minutes, whereas SAL-LDA takes almost 40 minutes to infer the spectral based latent semantics. This indicates SFF-FSTM is an efficient PTM compared with the classical non-sparse PTM such as SAL-LDA.

Number of training samples	Accuracy (%)	
	SFF-FSTM	SAL-LDA
80	94.55±1.02	88.33±1.82
60	93.10±1.42	87.00±1.17
40	89.76±1.09	84.35±1.46
20	82.92±1.42	77.23±1.62
10	78.23±1.28	71.46±1.58
5	75.65±1.56	66.29±2.09

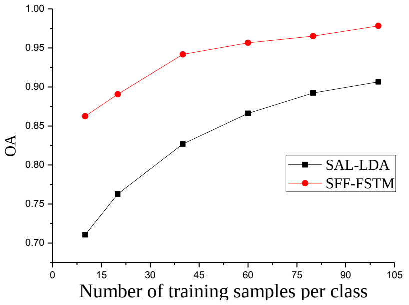
Table 5. Performance of SFF-FSTM and SAL-LDA for the UC Merced dataset with limited training samples

Number of training samples	Accuracy (%)	
	SFF-FSTM	SAL-LDA
100	97.83±0.93	90.65±1.05
80	96.52±0.77	89.23±0.99
60	95.65±0.82	86.62±1.02
40	94.17±0.89	82.69±0.96
20	89.07±1.04	76.29±1.21
10	86.26±1.13	71.06±1.45

Table 6. Performance of SFF-FSTM and SAL-LDA for the Google dataset of SIRI-WHU with limited training samples



(a)



(b)

Figure. 5. Classification accuracies with different numbers of training samples per class. (a) UC Merced dataset. (b) Google dataset of SIRI-WHU

4. CONCLUSION

In this paper, we have designed an effective and efficient approach—the semantic-feature fusion fully sparse topic model (SFF-FSTM)—for HSR imagery scene classification. The fully sparse topic model (FSTM) has been used for unsupervised dimension reduction of the large collection of documents first. By combining the novel use of FSTM and the semantic fusion of three distinctive features for HSR image scene classification, SFF-FSTM is able to presents a robust feature description for HSR imagery, and achieve comparative performance with limited training samples. The proposed SFF-FSTM can improve the performance of scene classification compared with other scene classification methods with the challenging UC Merced dataset and Google dataset of SIRI-WHU.

Nevertheless, image patches obtained by the uniform grid method might 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. In our future work, we plan to consider topic models which can take the correlation between image pairs into consideration.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China under Grant No. 41371344, State Key Laboratory of Earth Surface Processes and Resource Ecology under Grant No. 2015-KF-02, Program for Changjiang Scholars and Innovative Research Team in University under Grant No. IRT1278, Natural Science Foundation of Hubei Province under Grant No. 2015CFA002, and Open Research Fund Program of Shenzhen Key Laboratory of Spatial Smart Sensing and Services (Shenzhen University).

REFERENCES

Barla A., Odone F., and Verri A., 2003. Histogram intersection kernel for image classification. In: *Proceeding of International Conference on Image Processing*, Vol. 3, pp. III–513–16.

Bellens R., Gautama S., Martinez-Fonte L., Philips W., Chan J. C.-W., and Canters F., 2008. Improved classification of VHR images of urban areas using directional morphological profiles. *IEEE Trans. Geosci. Remote Sens.*, 46(10), pp. 2803–2813.

Blei D., Ng A., and Jordan M., 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res*., 3(5), pp. 993–1022.

Bosch A., Munoz X., and Marti R., 2007. Which is the best way to organize/classify images by content? *Image Vision Comput.*, 25(6), pp. 778–791.

Bosch A., Zisserman A., and Mu ñoz X., 2008. Scene Classification Using a Hybrid Generative/Discriminative Approach. *IEEE Trans. Pattern Anal. Mach. Intell*., 30(4), pp. 712–727.

Bratasanu D., Nedelcu I., and Datcu M., 2011. Bridging the semantic gap for satellite Image annotation and automatic mapping applications. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* (*JSTARS*), 4(1), pp. 193–204.

Chen S. and Tian Y., 2015. “Pyramid of Spatial Relatons for Scene-Level Land Use Classification,” *IEEE Trans. Geosci. Remote Sens.*, 53(4), pp. 1947–1957.

Cheriyadat A. M., 2014. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens*., 52(1), pp. 439–451.

Haralick R. M., Shanmugam K, and Dinstein I. H., 1973.Textural features for image classification. *IEEE Trans. Syst., Man, Cybern.*, SMC-3(6), pp. 610–621.

Hofmann T., 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn*., 42(1–2), pp. 177–196.

Huang K. and Aviyente S., 2008. Wavelet Feature Selection for Image Classification. *IEEE Transactions on Image Processing*, 17(9), pp. 1709-1720.

Lazebnik S., Schmid C., and Ponce J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *Proceeding of Conference on Computer Vision and Pattern Recognitio*n, Vol. 2, pp. 2169–2178.

Li énou M., Ma î re H., and Datcu M., 2010. Semantic annotation of satellite images using latent Dirichlet allocation. *IEEE Geosci. Remote Sens. Let*t., 7(1), pp. 28–32.

Li F.-F and Perona P., 2006. A Bayesian hierarchical model for learning natural scene categories. In: *Proceeding of IEEE Conference on Computer Vision and Pattern Recognitio*n, pp. 524–531.

Lowe D. G., 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2), pp. 91–110.

Mekhalfi M., Melgani F., Bazi Y., and Alajlan N., 2015. Land-Use Classification with Compressive Sensing Multifeature Fusion. IEEE Geosci. Remote Sens. Lett., vol. 12, no. 10, pp. 2155–2159, 2015.

Rizvi I. A. and Mohan B. K., 2011. 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.*, 49(12), pp. 4815–4820.

Shao W., Yang W., Xia G.-S., and Liu G., 2013. A Hierarchical Scheme of Multiple Feature Fusion for High-Resolution Satellite Scene Categorization. *Comput. Vision Syst*., pp. 324–333.

Shashanka M., Raj B., Smaragdis P, 2008. Sparse overcomplete latent variable decomposition of counts data. In: *Advances in neural information processing systems*, pp. 1313–1320.

Than K. and Ho T. B, 2012. Fully sparse topic models. In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (*ECML PKDD*), Bristol, UK. Vol. 7523 of LNCS, Springer, pp. 490-505.

Tilton J. C., Tarabalka Y., Montesano P. M., and Gofman E., 2012. Best merge region-growing segmentation with integrated nonadjacent region object aggregation. *IEEE Trans. Geosci. Remote Sens.*, 50(11), pp. 4454–4467.

Tokarczyk P., Wegner J. D., Walk S., and Schindler K., 2015. Features, Color Spaces, and Boosting: New Insights on Semantic Classification of Remote Sensing Images. *IEEE Trans. Geosci Remote Sens*., 53(1), pp. 280–295.

Văduva C., Gavăt I., and Datcu M., 2013. Latent Dirichlet allocation for spatial analysis of satellite Images. *IEEE Trans. Geosci. Remote Sens.*, 51(5), pp. 2770–2786.

Yang Y. and Newsam S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: *Proceedings of International Conference on ACM SIGSPATIAL GIS*, San Jose, California, pp. 270–279.

Zhao B., Zhong Y., and Zhang L., 2013. Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery. *Remote Sensing Letters*, 4(12), pp. 1204-1213.

Zhao B., Zhong Y., Xia G.-s., and Zhang L., 2016a. Dirichlet-Derived Multiple Topic Scene Classification Model Fusing Heterogeneous Features for High Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens*., 54(4), pp. 2108–2123.

Zhao B., Zhong Y., and Zhang L., 2016b. "A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 73-85.

Zhao B., Zhong Y., Zhang L., and Huang B., 2016c. The Fisher Kernel Coding Framework for High Spatial Resolution Scene Classification, *Remote Sensing*, vol. 8, no. 2, p. 157.

Zhao L.-J., Tang P., and Huo L.-Z., 2014. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sen*s., 7(12), pp. 4620–4631.

Zheng X., Sun X., Fu K., and Wang H., 2013. Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint. *IEEE Geosci. Remote Sens. Lett.* , 10(4), pp. 652–656.

Zhong Y., Zhu Q., and Zhang L., 2015. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.*, 53(11), pp. 6207–6222.

Zhu J., Xing, E.P, 2011. Sparse topical coding. In: *UAI*.