Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery

Qiqi Zhu, Yanfei Zhong, *Senior Member, IEEE*, Bei Zhao, Gui-Song Xia, *Member, IEEE*, and Liangpei Zhang, *Senior Member, IEEE*

***Abstract*—Scene classification has been studied to allow us to semantically interpret high spatial resolution (HSR) remote sensing imagery. The bag-of-visual-words (BOVW) model is an effective method for HSR image scene classification. However, the traditional BOVW model only captures the local patterns of images by utilizing local features. In this letter, a local–global feature bag-of-visual-words scene classifier (LGFBOVW) is proposed for HSR imagery. In LGFBOVW, the shape-based invariant texture index is designed as the global texture feature, the mean and standard deviation values are employed as the local spectral feature, and the dense scale-invariant feature transform (SIFT) feature is employed as the structural feature. The LGFBOVW can effectively combine the local and global features by an appropriate feature fusion strategy at histogram level. Experimental results on UC Merced and Google data sets of SIRI-WHU demonstrate that the proposed method outperforms the state-of-the-art scene classification methods for HSR imagery.**

***Index Terms*—Bag-of-visual-words (BOVW), high spatial resolution (HSR), local and global features, remote sensing, scene classification.**

I. Introduction

WITH the development of modern satellite technologies, high spatial resolution (HSR) remote sensing images can provide detailed spatial information. Object-based and contextual-based methods are both used for precise object recognition [1], [21]. Due to the semantic phenomena known as "synonymy" and "homonymy," these methods usually cannot capture the complex semantic information. This can be described as the divergence between the "information" that is derived from the data and the "knowledge" specific to each user and application, namely, the "semantic gap" [2].

Scene classification methods, which can automatically label an image from a set of semantic categories [3], have been proposed to obtain the semantic information from HSR imagery [4]–[6]. The bag-of-visual-words (BOVW) model [24] is a classical intermediate feature representation method that can be used to bridge the semantic gap [7]. In general, a single feature is utilized to describe the images, which is inadequate [22]. BOVW-based multiple local feature scene classification methods have also been developed [9]. However, the spectral or scale-invariant feature transform (SIFT) features can only describe the local information, which is continuous, and they do not contain features captured from the global perspective. Moreover, the visual words acquired by clustering the long feature vector using k-means clustering are not adequate [10]. The fact that there are a lot of mixed pixels in HSR imagery is not in accordance with the hard-assignment-based clustering algorithm of k-means clustering [11].

In this letter, a local–global feature BOVW scene classifier (LGFBOVW) is proposed for HSR imagery. The main contributions of this letter are presented as follows.

In LGFBOVW, spectral and SIFT features are considered as the local feature, and a new texture feature is designed as the global feature. By capturing the characteristics of HSR imagery from both the local and global, discrete and continuous perspective, LGFBOVW presents a robust feature description. For such whole-image categorization tasks, global information which is complementary from the local features should be considered to comprehensively understand the semantics of the scenes. The shape-based invariant texture index (SITI) is designed as the global texture feature. In contrast to the popular SIFT feature, which focuses more on the corner and edge attributes, SITI captures more of the shape characteristics of a scene. The fusion of the three features at histogram level is able to circumvent the inadequate fusion capacity of the k-means algorithm, and the mutual interference between different features. The incorporation of support vector machine (SVM) with a histogram intersection kernel (HIK) is effective in increasing the discrimination of different scenes. Experimental results on the UC Merced and Google datasets demonstrate that the proposed LGFBOVW produces a commendable performance.

The remainder of this letter is organized as follows. Section II describes the details of the proposed LGFBOVW for HSR imagery scene classification. A description of the data sets and an analysis of the experimental results are presented in Section III. Finally, the conclusions are drawn in Section IV.

Manuscript received September 27, 2015; revised December 5, 2015; accepted December 18, 2015. Date of publication May 6, 2016; date of current version May 19, 2016. This work was supported by the National Natural Science Foundation of China under Grant 41371344, by the State Key Laboratory of Earth Surface Processes and Resource Ecology under Grant 2015-KF-02, and by the Open Research Fund Program of Shenzhen Key Laboratory of Spatial Smart Sensing and Services (Shenzhen University). *(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 (e-mail: zhongyanfei@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/LGRS.2015.2513443

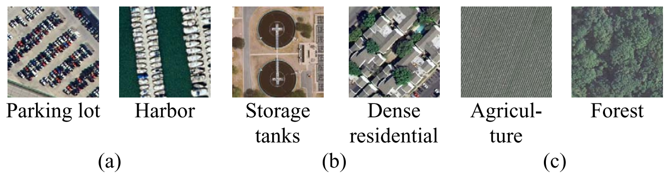

Fig. 1. HSR images of the parking lot, harbor, storage tanks, dense residential, forest, and agriculture scene classes: (a) importance of the spectral characteristics for HSR images; (b) importance of the structural characteristics for HSR images; and (c) importance of the textural characteristics for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

II. Background

A. Local and Global Feature Extraction Based on the BOVW Model for HSR Imagery Scene Classification

In Fig. 1(a), it is difficult to distinguish parking lot from harbor, from both the structural and textural characteristics. 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 mainly differ in structural characteristics. The forest and agriculture scenes are similar in spectral and structural characteristics in Fig. 1(c), but they differ greatly in the textural information from the global perspective. Hence, two local features and a global feature are designed for the HSR imagery scene classification task.

A uniform grid method has been proved to be more effective than other sampling methods such as random sampling [12]. To extract the local spectral and SIFT features, the HSR image is first partitioned into patches utilizing a uniform grid method. The feature descriptors are then extracted from the local patches. The k-means clustering is used to quantize the descriptors into a codebook of visual words [23]. With the visual words representing the specific local pattern, we have a codebook describing all the kinds of local image patterns. In contrast, the global feature is captured based on the whole image, and it is a compact and discriminative representation of the HSR image. SITI, as the global discrete feature, can be directly extracted to acquire the global feature vectors.

The first- and second-order statistics (the mean and standard deviation values) of the patches are computed in each spectral channel as the spectral feature. We let n be the number of pixels in the sampled patch, and v_{ij} denotes the jth band value of the ith pixel in a patch. The mean (mean_j) and standard deviation (std_j) of the patch are calculated according to

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

The gray dense SIFT descriptor [13] with 128 dimensions is extracted as the structural feature. This was inspired by previous work, in which dense features performed better for scene classification [12].

SITI [14], which was first proposed for texture image retrieval and classification, is employed as the texture feature. SITI is based on the topographic map of images, which is the complete set of level lines of the images. This is different from the SIFT feature, which is the edge-based scene characteristics captured from the local keypoints. SITI focuses more on the global shape characteristics, such as the elongation and the compactness of the shape. Here, the shapes, which are the interiors of the connected components of the level lines, are the basic elements on which SITI is performed. The elongation histogram (EH), the compactness histogram (CpH), the scale ratio histogram (SRH), and the contrast histogram (CtH) are combined as the feature descriptor for HSR imagery, which is defined as the "similarity invariant local features" (SI) [14]. EH, CpH, SRH, and CtH of the SITI vector are calculated according to (2) and (3), where λ_1 and λ_2 are the eigenvalues of the inertia matrix, $\mu_{00}(s)$ is the area of the shape s, $\langle \bullet \rangle_{s' \in M}$ is the mean operator on M, and M is the partial neighborhood of order M of s. At each pixel x, $s(x)$ is the smallest shape of the topographic map containing x, and $\mathrm{mean}_{s(x)}(u)$ and $\mathrm{var}_{s(x)}(u)$ are, respectively, the mean and variance of u over $s(x)$. That is

$$\text{Elongation} : \in = \frac{\lambda_2}{\lambda_1}, \ \text{Compactness} : \kappa = \frac{1}{4\pi\sqrt{\lambda_1 \lambda_2}} \tag{2}$$

$$\text{Scale ratio} : \ \alpha(s) = \frac{\mu_{00}(s)}{\langle \mu_{00}(s') \rangle_{s' \in \mathbb{N}^M}}$$

$$\text{Contrast} : \ \gamma(x) = \frac{\mu(x) - \mathrm{mean}_{s(x)}(u)}{\sqrt{\mathrm{var}_{s(x)}(u)}}. \tag{3}$$

The parameters in our experiments with the SITI feature were set according to the recommendation of the author [14].

B. Local and Global Feature Fusion Based on the BOVW Model for HSR Imagery Scene Classification

On acquiring the feature vectors, the same visual word may have different feature values due to the scale, rotation, and illumination variation of the image scenes. Hence, the spectral feature descriptor is then quantized by k-means clustering, and the image patches with similar feature values can be clustered to the same visual word. Suppose that we have a collection of N images $D = \{d_1, \ldots, d_N\}$, where an image is split into P patches. For each patch, the spectral feature descriptors $X = \{x_1, \ldots, x_i, \ldots, x_P\}$ are extracted and quantized according to (4). K cluster centers C are randomly initialized, and S_j denotes the descriptors belonging to cluster center C_j. By minimizing the sum of squared Euclidean distances between descriptors x_i and their nearest cluster center C_j, the visual words for an image are acquired. That is

$$D(X, C) = \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - C_j\|. \tag{4}$$

A statistical analysis of the frequency of each visual word is performed and a 1-D histogram with V_1 bins is generated. Similarly, the SIFT feature is quantized into a 1-D histogram with V_2 bins by k-means clustering. Here, V_1 or V_2 represents the number of visual words for the local continuous features. Except for the local continuous features, there are three other types of features. For the local discrete features, the frequency of the same discrete values can be counted to represent the images. The global discrete feature can be directly used as the 1-D histogram of the scenes, and the global continuous

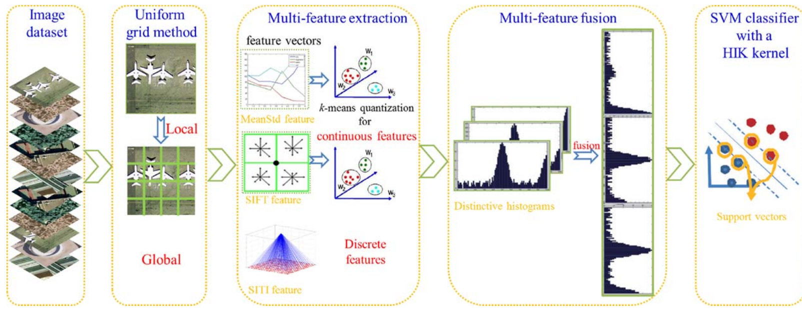

Fig. 2. Proposed HSR scene classification based on LGFBOVW.

TABLE I
CLASSIFICATION ACCURACY (%) FOR THE UC MERCED DATA SET WITH THE DIFFERENT METHODS

Method \ Feature	Spectral	Structure	Texture	
	SPECTRAL	SITI	GLCM	SIFT
BOVW	85.30±1.67	81.54±1.46	78.27±1.68	87.31±1.40
BOVW-SPE-SI	86.38±1.38			
LGFBOVW-FV	88.02±1.87			
LGFBOVW-Li	88.33±1.56			
LGFBOVW	**96.88±1.32**			

TABLE II
COMPARISON WITH THE EXPERIMENTAL RESULTS OF PREVIOUS METHODS FOR THE UC MERCED DATA SET

Methods	SPM [16]	Yang and Newsam [4]	Cheriyadat [5]	Chen and Tian [8]	Mekhalfi *et al.* [17]	LGFBOVW
Accuracy (%)	82.30±1.48	81.19	81.67±1.23	89.10	94.33	**96.88±1.32**

feature can be stretched into a 1-D histogram with a certain scale. SITI is adopted for the global feature extraction. SITI, as the global discrete feature, can be extracted and used to directly represent the 1-D histogram of the HSR image scene. Supposing that there are M images with a texture histogram of V_3 bins, then three histograms are fused at histogram level to generate an LGFBOVW representation with $(V_1 + V_2 + V_3) \times M$ dimensions for all the images.

After obtaining the LGFBOVW representation, the final classification step utilizes the SVM classifier with a HIK to predict the scene label. The HIK measures the degree of similarity between two histograms, to deal with the scale changes. We let $\tilde{\mathbf{V}} = (\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_M)$ be the LGFBOVW representation vectors of M images, and the HIK is calculated according to (5). The HIK has been successfully applied in image classification using color histogram features [15]. With the generated LGFBOVW representation as an extended histogram, SVM with a HIK is able to enlarge the discrimination of LGFBOVW representation vector. Scene classification based on LGFBOVW is shown in Fig. 2. That

$$K(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j) = \sum_k \min(\tilde{\mathbf{v}}_{i,k}, \tilde{\mathbf{v}}_{j,k})^2. \tag{5}$$

III. EXPERIMENTS AND ANALYSIS

A. Experimental Design

A 12-class Google data set of SIRI-WHU and the commonly used 21-class UC Merced data set were evaluated to test the performance of LGFBOVW. In Tables I and II, SPECTRAL, SITI, and SIFT denote scene classification utilizing the spectral feature based on the mean and standard deviation, the SITI-based texture feature, and the SIFT-based structural feature, respectively. BOVW-SP-SIFT denotes BOVW employing the spectral and SIFT feature. LGFBOVW-FV is BOVW combining three features at the feature vector level. LGFBOVW-Li represents LGFBOVW utilizing SVM with a linear kernel. To further evaluate the performance of LGFBOVW, the experimental results utilizing spatial pyramid matching (SPM) [16] and the experimental results on the UC Merced data set, as published in the latest papers by Yang and Newsam in 2010 [4], Cheriyadat in 2014 [5], Chen and Tian in 2015 [8], and Mekhalfi *et al.* in 2015 [17], are shown for comparison. SPM employed the dense gray SIFT, and the spatial pyramid layer was optimally selected as one. The experimental results on the Google data set of the SIRI-WHU data set utilizing the methods of latent dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA), as published by Lienou *et al.* [18] and Bosch *et al.* [19], respectively, are also shown for comparison.

In the experiments, the images were uniformly sampled with a patch size and spacing of 8 and 4 pixels, respectively, for SPECTRAL with the two data sets. The patch size and spacing for SIFT were set to 16 and 8 for the UC Merced data set, and 8 and 4 for the Google data set of SIRI-WHU, respectively. Considering the accuracy and the efficiency, the number of visual words for SPECTRAL and SIFT were optimally set to 1000, with 200 for SITI. To test the stability of the proposed LGFBOVW, the different methods were executed 100 times by a random selection of training samples.

B. Experiment 1: The UC Merced Image Data Set

The UC Merced data set was downloaded from the USGS National Map Urban Area Imagery collection [4]. This data set consists of 21 land-use scenes (see 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 in [4], 80 samples were randomly selected per class from the UC Merced data set for training, and the rest were kept for testing.

Fig. 3. Example images from the 21-class UC Merced dataset.

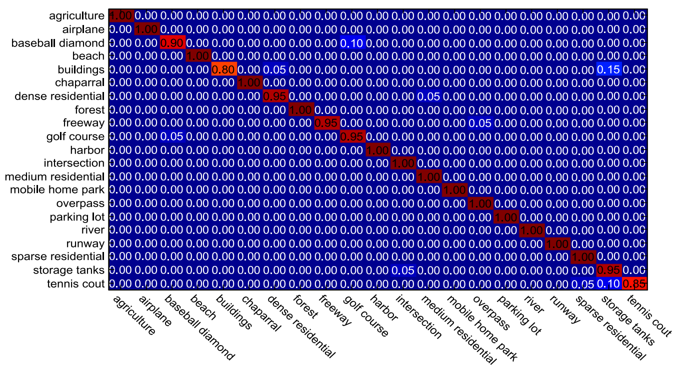

Fig. 4. Confusion matrix for LGFBOVW with the UC Merced data set.

The classification performances of the different methods with the UC Merced data set are reported in Table I. As shown in Table I, the classification result of SITI, i.e., 81.54% $\pm$ 1.46% for the BOVW model, is better than that of the gray level co-occurrence matrix (GLCM) for BOVM at 78.27% $\pm$ 1.68%. The classification performance of LGFBOVW is the best among all. LGFBOVW outperforms BOVW-SPE-SI, which indicates that SITI is a complementary feature for the spectral and SIFT features. Comparing LGFBOVW and LGFBOVW-FV, it can be seen that the fusion of the features at histogram level is more appropriate. The comparison of LGFBOVW and LGFBOVW-Li shows that SVM with a HIK kernel is able to improve the discriminative ability. In Table II, it is shown that the proposed LGFBOVW is superior to the performance of SPM [16], the Yang and Newsam method [4], the Cheriyadat method [5], the Chen and Tian method [8], and the Mekhalfi *et al.* method [17], and it exceeds the state-of-the-art performance with the UC Merced data set. This confirms that LGFBOVW is a competitive method for HSR imagery. Fig. 4 displays the confusion matrix of LGFBOVW for the 21-class UC Merced data set. As can be seen in the confusion matrix, 14 scene classes can be fully recognized by LGFBOVW, except for the baseball diamond, buildings, dense residen, freeway, golf course, storage tanks, and tennis court classes.

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

The Google data set was acquired from Google Earth (Google Inc.), mainly covering urban areas in China, and the scene image data set is designed by the Intelligent Data Extraction and Analysis of Remote Sensing (RS_IDEA) Group in Wuhan University (SIRI-WHU) [20]. It consists of 12 land-use classes, which are labeled as follows: agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water, as shown in Fig. 5. Each class separately contains 200 images, which were cropped to 200 $\times$ 200 pixels, with a spatial resolution of 2 m. The 100 training samples were randomly selected per class, and the remaining samples were retained for testing.

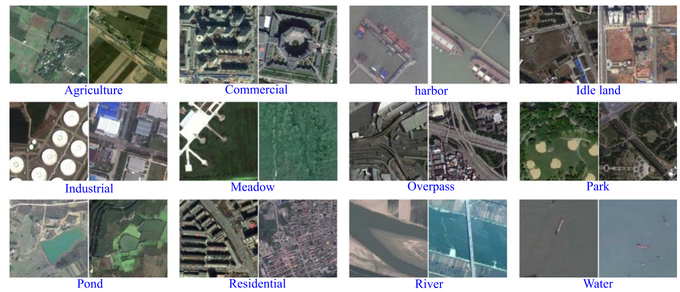

Fig. 5. Example images from the Google dataset of SIRI-WHU.

TABLE III
CLASSIFICATION ACCURACIES (%) FOR THE GOOGLE DATASET OF SIRI-WHU WITH DIFFERENT METHODS

Feature / Method	Spectral	Structure	Texture	
	SPECTRAL	SIFT	GLCM	SITI
BOVW	86.51±0.92	82.14±0.90	75.83±1.65	79.23±1.01
BOVW-SPE-SI	89.84±0.96			
LGFBOVW-FV	91.68±1.27			
LGFBOVW-Li	92.17±1.21			
LGFBOVW	**96.96±0.95**			

TABLE IV
COMPARISON WITH THE EXPERIMENTAL RESULTS OF PREVIOUS METHODS FOR THE GOOGLE DATASET OF SIRI-WHU

Methods	SPM [16]	LDA [18]	PLSA [19]	LGFBOVW
Accuracy (%)	77.69±1.01	60.32±1.20	89.60±0.89	**96.96±0.95**

The classification performances of different methods with the Google data set of SIRI-WHU are reported in Table III. As shown in Table III, the BOVW employing the SITI-based feature achieves a better performance than it does with GLCM-based feature. This indicates that SITI can capture more texture details than GLCM. The result of LGFBOVW is the best among all. The comparison of the results between BOVW-SPE-SI and LGFBOVW infers that the novel introduction of SITI to HSR imagery scene classification is effective. In addition, LGFBOVW is superior to LGFBOVW-Li, which proves that the simple HIK kernel outperforms the commonly used linear kernel. In Table IV, compared to the other methods, the highest accuracy is acquired by the proposed LGFBOVW.

D. Sensitivity Analysis

To study the effect of the number of visual words for the different feature-based BOVW models with the UC Merced data set, the patch scale and spacing were kept at 8 and 4 for the spectral feature, and 16 and 8 for the SIFT feature, respectively. The number of visual words for the spectral and SIFT features was then varied over the range of [500, 1000, 1500, 2000, 2500, 3000], and [120, 160, 200, 240, 280, 320] for the SITI

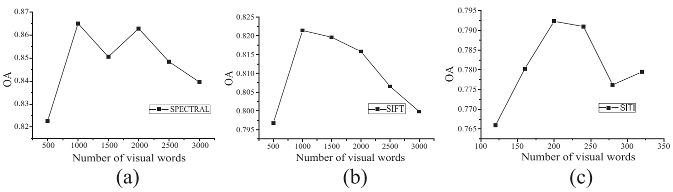

Fig. 6. Sensitivity analysis. (a) SPECTRAL. (b) SIFT. (c) SITI.

feature. In Fig. 6, it is shown that the results of the three feature-based BOVW models display a great fluctuation in relation to the number of visual words. SPECTRAL obtains the highest accuracy of 86.51% when the number of visual words is 1000, while SIFT obtains an accuracy of 82.14% when the number of visual words is 1000. SITI provides an accuracy of 79.23% when the number of visual words is 200.

IV. Conclusion

In this letter, we have designed a simple but effective approach (LGFBOVW) for HSR imagery scene classification. By introducing the novel use of the SITI for the texture feature of the BOVW-based HSR imagery scene classification, LGFBOVW captures HSR imagery, from both the global and local perspective, and presents a robust feature description for HSR imagery. The LGFBOVW outperforms the state-of-the-art performance with the UC Merced data set, and it provides novel feature strategies for UC Merced data set classification utilizing the BOVW model.

Nevertheless, the experiments with the UC Merced data set and Google data set of SIRI-WHU were conducted by selecting 80% and 50% of the samples as training samples, respectively. Reducing the number of training samples would be more practical. In addition, 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 may be considered. In our future work, we plan to consider models which can relax the normalization constraints of the probabilistic topic model, and topic models which consider the correlation between image pairs.

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] 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.

[3] A. Bosch, X. Munoz, and R. Marti, "Which is the best way to organize/classify images by content?" *Image Vis. Comput.*, vol. 25, pp. 778–791, Jul. 2007.

[4] 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.

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

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

[7] B. Zhao, Y. Zhong, L. Zhang, and B. Huang, "The Fisher kernel coding framework for high spatial resolution scene classification," *Remote Sens.*, vol. 8, no. 2, p. 157, 2016, doi: 10.3390/rs8020157.

[8] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, 2015.

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

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

[11] S. P. Lloyd, "Least square quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1957.

[12] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 524–531.

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

[14] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 382–403, Jul. 2010.

[15] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Proc. IEEE Int. Conf. Image Process.*, 2003, vol. 3, pp. III-513–III-16.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, vol. 2, pp. 2169–2178.

[17] M. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "Land-use classification with compressive sensing multifeature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2155–2159, 2015.

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

[19] 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.

[20] B. Zhao, Y. Zhong, G.-s. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model fusing heterogeneous features for high resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[21] I. A. Rizvi and B. K. Mohan, "Object-based image analysis of high-resolution satellite images using modified cloud basis function neural network and probabilistic relaxation labeling process," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4815–4820, Dec. 2011.

[22] H. Sridharan and A. Cheriyadat, "Bag of Lines (BoL) for improved aerial scene representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 676–680, Mar. 2015.

[23] L. Weizman and J. Goldberger, "Urban-area segmentation using visual words," *Remote Sens. Lett.*, vol. 6, no. 3, pp. 388–392, 2009.

[24] G. Csurka *et al.*, "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop Statist. Learn. Comput. Vis.*, Prague, Czech Republic, 2004.