APPLICATIONS PAPER

Tea Garden Detection from High-Resolution Imagery Using a Scene-Based Framework

Xin Huang, Zerun Zhu, Yansheng Li, and Bo Wu, and Michael Yang

Abstract

Tea cultivation has a long history in China, and it is one of the pillar industries of the Chinese agricultural economy. It is therefore necessary to map tea gardens for their ongoing management. However, the previous studies have relied on fieldwork to achieve this task, which is time-consuming. In this paper, we propose a framework to map tea gardens using high-resolution remotely sensed imagery, including three scene-based methods: the bag-of-visual-words (BOVW) model, supervised latent Dirichlet allocation (sLDA), and the unsupervised convolutional neural network (UCNN). These methods can develop direct and holistic semantic representations for tea garden scenes composed of multiple sub-objects, thus they are more suitable than the traditional pixel-based or object-based methods, which focus on the local characteristics of pixels or objects. In the experiments undertaken in this study, the three different methods were tested on four datasets from Longyan (Oolong tea), Hangzhou (Longjing tea), and Puer (Puer tea). All the methods achieved a good performance, both quantitatively and visually, and the UCNN outperformed the other methods. Moreover, it was found that the addition of textural features improved the accuracy of the BOVW and sLDA models, but had no effect on the UCNN.

Introduction

Tea is one of the most famous beverages in the world, and the consumption of tea is growing faster than that of coffee and cocoa (Cabrera *et al.*, 2006). China was not only the first country to cultivate tea, but it is also one of the main producing countries (Dutta *et al.*, 2010). The cultivation and production of tea plays an important part in Chinese agriculture, and has a significant impact on the economic development of rural areas. Tea is produced in most provinces of southern China and is the major cash crop for many villages and towns. Thus, it is necessary to monitor and assess the tea gardens. However, this task is usually achieved by fieldwork, which is labor- and time-intensive. Meanwhile, remote sensing technology has been widely employed to detect and map crops, such as sugarcane (Vieira *et al.*, 2012), paddy rice (Qin *et al.*, 2015), and coca (Pesaresi, 2008). Although it is an effective and convenient tool, few studies have been so far carried out for tea garden mapping using remotely sensed imagery.

The tea plant, an evergreen bush, can be easily confused with other woody vegetation in spectral characteristics. However, tea gardens exhibit unique textural features, which can be used to distinguish them from other vegetation, due to the unique form of tea cultivation. In high-resolution optical imagery, tea gardens are composed of multiple object types with a certain spatial and structural pattern: (1) tea bushes are generally planted in rows, showing obvious gaps between the rows; and (2) tea gardens contain not only the tea plants, but also bare soil, individual trees, and shadows (see Figure 1). In this regard, a tea garden can be considered as a semantic scene consisting of multiple interrelated sub-objects, rather than a single land-cover type. Because the pixel-based (Damodaran *et al.*, 2017; Huang *et al.*, 2014) or object-based (Ming *et al.*, 2016; Zhu *et al.*, 2017) image analysis methods, modeling the scene in a bottom-up manner, have difficulty in obtaining the holistic semantic representation of a scene, they are not suitable for detecting tea garden scenes with complex sub-categories and spatial patterns. On the other hand, scene-based analysis techniques have been proved to be a more productive approach for the interpretation of high-resolution remotely sensed imagery (Cheriyadat, 2014; Huang *et al.*, 2015). Therefore, in this study, we propose a series of scene-based interpretation models with spectral-textural features to detect tea gardens from high-resolution imagery.

In recent years, various scene classification approaches have been proposed. One particularly efficient method, the bag-of-visual-words (BOVW) model (Sivic and Zisserman, 2003), has been widely used for remote sensing semantic classification (Cheriyadat, 2014; Sheng *et al.*, 2012). In the classic BOVW model, an image is represented by a set of visual words, which are generated by clustering the local patches. Subsequently, topic models, such as probabilistic latent semantic analysis (pLSA) (Hofmann, 2001) and latent Dirichlet allocation (LDA) (Blei *et al.*, 2003), have been adopted to extract the latent semantic topic features from the BOVW representation and classify the scenes with the semantic topic features. The pLSA model uses a probabilistic approach to model an image, representing the frequency of the visual words as a finite mixture of an intermediate set of topics. LDA improves on pLSA by introducing a Dirichlet distribution into the topic mixture, thus overcoming the overfitting problem of pLSA. More recently, the sLDA model (Jon and David, 2008), which extends LDA by adding a response variable to indicate the class of the scenes in a generative process, has been proposed and successfully utilized in image annotation and satellite

Xin Huang is with the School of Remote Sensing Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; and the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (xhuang@whu.edu.cn).

Zerun Zhu is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China.

Yansheng Li is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China.

Bo Wu is with the School of Geography and Environment, Jiangxi Normal University, Nanchang 330022, China.

Michael Yang is with the Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede 7500 AE, The Netherlands.

Photogrammetric Engineering & Remote Sensing
Vol. 84, No. 11, November 2018, pp. 723–731.
0099-1112/18/723–731

doi: 10.14358/PERS.84.11.723

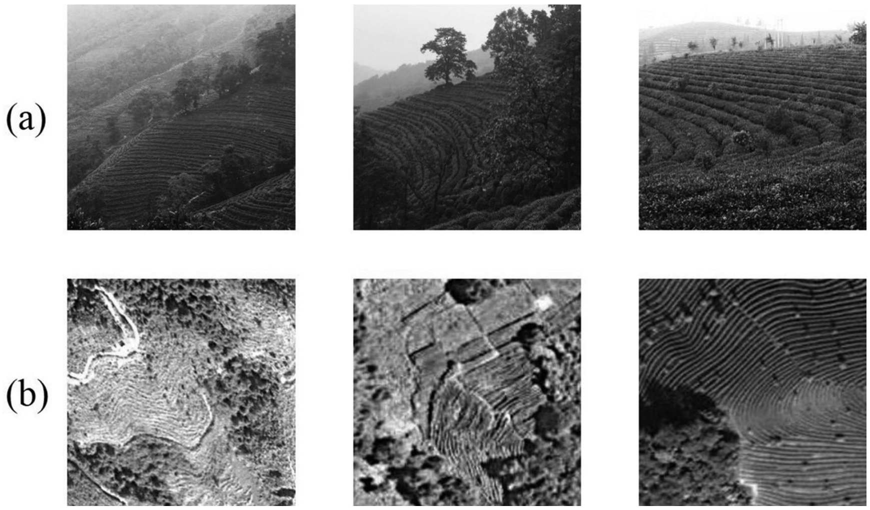


Figure 1. Examples of tea gardens: (a) Digital camera photographs, and (b) Google Earth images.

scene classification (Chong *et al.*, 2009; Putthividhy *et al.*, 2010; Huang *et al.*, 2015).

However, the BOVW and topic models rely on the establishment of empirically designed features to depict the local patches of the images as visual words. In order to overcome this limitation, an increasing amount of research has focused on unsupervised machine learning methods to autonomously extract adaptive and suitable features from unlabeled input data. For example, Coates *et al.* (2010) built a single-layer UCNN for unsupervised feature learning. They used different unsupervised learning algorithms to generate local convolutional features (i.e., function bases), and found that *k*-means clustering, which is an extremely simple learning algorithm, achieved the best performance. Blum *et al.* (2012) applied the network proposed in Coates *et al.* (2010) to object recognition from natural images with depth information. Dosovitskiy *et al.* (2014) developed a multi-layer UCNN to learn feature representations from unlabeled images, and the learned features performed well in natural image classification. Recently, in order to classify remotely sensed scenes, Li *et al.* (2016) trained a multi-layer UCNN using *k*-means clustering to autonomously mine complex structure features from high-resolution images, and used support vector machine (SVM) for the final scene classification. In this study, the features extracted by the UCNN achieved a better scene classification accuracy than BOVW and sparse coding.

To the best of our knowledge, little research has been so performed concerning tea garden detection from remote sensing data. However, this is necessary since tea cultivation plays an important part in Chinese agriculture, but the current tea garden monitoring relies on field investigation, which is time-consuming and labor-intensive. In this context, in the proposed scene-based framework, high-resolution satellite images are employed to detect tea gardens, since these images can provide abundant spatial and textural information. Considering that a tea garden is a semantic scene composed of a variety of interrelated objects in a high-resolution image, we propose to apply scene-based semantic learning methods for tea garden detection, including the following experimental configurations: (1) BOVW is used to represent the scenes with spectral and Gabor textural features. An SVM classifier is then employed to classify the representation into tea gardens and non-tea gardens; (2) sLDA is used to extract the topic features from the BOVW representation of the scenes and predict the category label of each scene; and 3) A multi-layer UCNN is trained to generate discriminative features from the original spectral images, and the derived features are also classified by SVM.

The rest of this paper is organized as follows. The next section introduces the tea garden detection framework, followed by a description of the datasets and the experimental setup. The next section presents the detection results and discussion with the different methods and features. The last section concludes the paper.

Methodology

In this section, we introduce the scene classification methods employed in this study (i.e., BOVW, sLDA, and the UCNN). Subsequently, the proposed scene-based tea garden detection framework is described in detail.

Topic Scene Classification Models

BOVW is the basis of the topic models, and thus it is presented before introducing the sLDA model. The BOVW model was derived from a text analysis method which represents a document by the word frequencies, ignoring their order. The idea was then applied to images by utilizing the visual words formed by vector quantizing the visual features. The BOVW representation is constructed in two stages, as shown in Figure 2, i.e., visual word learning and feature encoding. During the visual word learning, the remotely sensed images are divided into patches, and the spectral or textural information of these patches is extracted to generate feature vectors which can describe the patches. We then quantify the spectral and textural descriptors using the k-means clustering algorithm. The cluster centers, which are known as “visual words”, form a dictionary. In the feature encoding, an unrepresented scene is split into several patches. Each patch is assigned to the label of the closest cluster center after extracting features of the patch. In this way, an image can be represented by a frequency histogram of the labeled patches. The histogram can be regarded as a feature vector for the subsequent classification, whose size is equal to the size of the dictionary.

The BOVW model represents a scene as a text document by the frequencies of the visual words. The LDA model (Blei *et al.*, 2003), which is a generative probabilistic model from the statistical text literature, characterizes the scene as random mixtures over latent topics, where each topic in turn is described by a distribution over the visual words in the dictionary. The process of LDA to generate a scene *d* can then be described as follows (as shown in Figure 3):

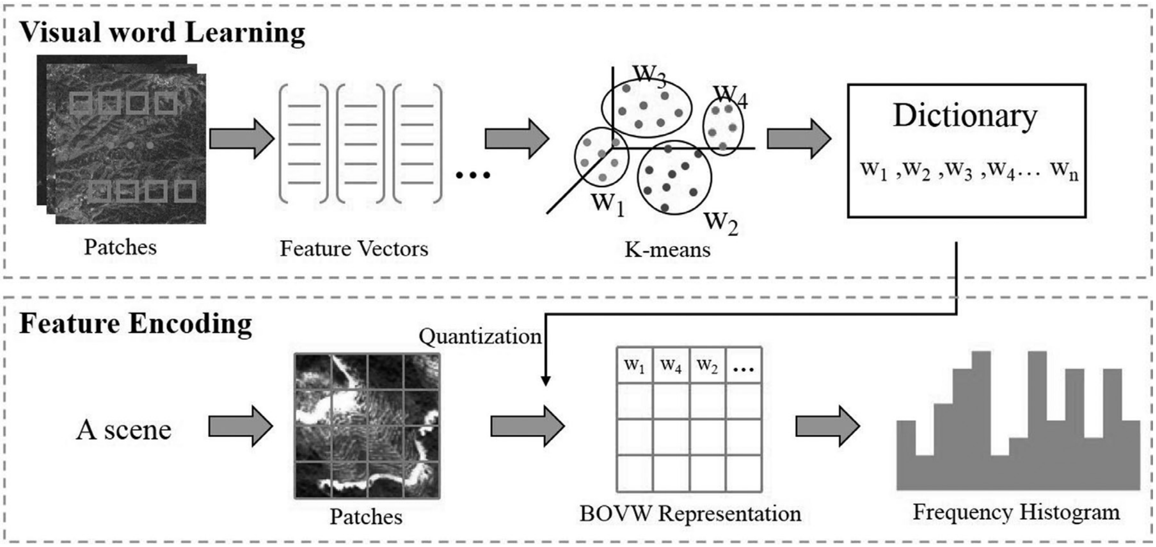


Figure 2. Flowchart of the BOVW model.

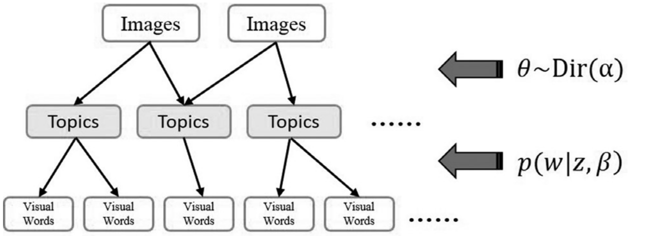


Figure 3. Generative process of LDA.

1. For the scene d, a K-dimensional topic proportion θ is chosen according to a Dirichlet distribution Dir(α), where K is the number of topics.
2. For each visual word position in the scene, a topic z is first chosen from the multinomial distribution Mul(θ), and then a visual word w is chosen from $p(w \mid z, \beta)$, a multinomial probability conditioned on topic z.

The above process shows that the model is controlled by α and β, and thus in the learning stage, our goal is to find the two parameters such that the log likelihood of the image dataset is maximized. It is clear that LDA is an unsupervised model, and the estimated topics are not specifically for classification. To mark the category of a document directly, Jon and David (2008) developed sLDA, which is a supervised variant of LDA, and proved that sLDA fitted the category of documents better than LDA. Since we are more concerned about the category than the topics of the scene, the sLDA model is applied to tea garden scene detection in the proposed framework. As described in Jon and David (2008), sLDA adds a response variable which denotes the categories of the scenes (i.e., tea gardens and non-tea gardens in our study) in the generative process of LDA. After generating a scene, the response variable associated with this scene is also generated. Thus, the learned model can then be used to classify the unknown scenes.

Unsupervised Feature Learning

In the proposed framework, the UCNN, via the plain k-means clustering method, is constructed to achieve unsupervised multi-layer feature learning (Li *et al.*, 2016). As depicted in Figure 4, the UCNN is composed of two feature extraction layers, and each layer contains three operations: convolution, local pooling, and global pooling. In the following, we take the first feature extraction layer as an example to introduce the three operations.

1. Convolution operation: The function of the convolution operation is feature mapping, which is defined under the constraint of function bases. The function bases need to be generated by an unsupervised learning algorithm, and k-means clustering is utilized in the proposed framework due to its good performance (Coates et al., 2010). As described in Li et al. (2016), the unlabeled patches with dimension $w - w - d$ are randomly sampled from the original image scenes, where w denotes the size of the receptive field, and d is the number of image channels. We can then construct the feature set $\mathrm{X} = \{x^1, x^2, \ldots, x^M\}$, where $x^i \in R^N (N = w * w * d)$ denotes the vectorization vector of the $i - th$ patch. After preprocessing by intensity normalization and zero component analysis whitening, the feature set X is clustered by the k-means clustering approach, and the clustering centers form the function bases $\mathrm{C} = \{c^1, c^2, \ldots, c^K\}$, with $c^i \in R^K$. Once the function bases are generated, the convolution operation can be defined as follows. Let p denote the vectorization vector of one sliding patch in the input image I, then this patch can be mapped onto the sparse feature vector $f \in R^K$:

$$f_k = \max\{0, \mu(z) - z_k\} \tag{1}$$

where $z_k = \| p - c^K \|_2$, $k = 1, 2, \ldots, K$ and $\mu(z)$ is the mean of the elements of z. Through the convolution operation, we produce the feature map F of the input image I with dimension $(h - n) - (h - n) - d$, where h denotes the size of image I.
2. Local pooling operation: This pooling operation is implemented to keep slight translation and rotation invariance. Here, the local pooling operation is defined as:

a. $$\mathrm{L}(i/s, j/s, k) = \max(\mathrm{F}(i - s/2 : i + s/2 : j - s/2 : j + s/2 :)) \tag{2}$$

where $k = 1, 2, \ldots, K$ and s denotes the local window size of the pooling operation,
3. Global pooling operation: The aim of the global pooling operation is to reduce the dimension of the feature. In the implementation, the output of the local pooling operation

is divided into four quarters, and the averaging operation is employed for each quarter. We let g^1, g^2, g^3, g^4 denote the corresponding feature vectors of every quarter, and then the global pooling result can be expressed as $G = \{g^1, g^2, g^3, g^4\}$, where the dimension of G is $4*K$.

The multi-layer features can be obtained by repeating the convolution, local pooling, and global pooling operations. Specifically, taking the original image as the input of the first layer and implementing the three operations, we can obtain the feature extracted by the first layer, i.e., the result of the global pooling operation. Then, taking the result of the local pooling operation in the first layer as the input of the second layer, we can obtain the feature of the second layer by the three operations. The features of the higher layers can be similarly extracted. Finally, we integrate the features of each layer as the final multi-layer features.

The Proposed Tea Garden Scene Detection Framework

The proposed scene-based framework for tea garden detection is illustrated in Figure 5. First, the whole remotely sensed image is divided into a series of scenes, with each scene referring to an image block with a certain semantic category. Since a tea garden does not always occupy all of a scene, we utilize the overlapping regions to decrease the omissions, and the size of an overlapping region is half of the scene. Then, based on the extracted spectral and textural features, the BOVW, sLDA, and UCNN representations of each scene are computed, respectively. Note that the dictionary in BOVW or the UCNN is trained by unlabeled data before detection. Finally, SVM, which is an efficient classifier for remotely sensed image classification (Huang and Zhang, 2013; Qin, 2015), is trained using a small number of labeled samples to classify these representations into tea gardens and non-tea gardens, expect for the sLDA model, which can directly predict the category of each scene after training. In addition, for the purpose of a comparison, the scenes with spectral and spectral-textural features are also directly classified with SVM.

In our work, the Gabor filter (Lee, 1996) is used to extract the textural features due to its high efficacy in remote sensing image texture description (Reis and Taşdemir, 2011; Wang *et al.*, 2014). It is defined as:

$$G(x,y) = \frac{1}{2\pi ab} e^{-\pi(\frac{x^2}{a^2}+\frac{y^2}{b^2})} e^{i(ux+vy)} \qquad (3)$$

where a,b denote the scale along x and y, respectively, and u,v are the spatial frequencies of the filter in the frequency domain. To improve the calculation efficiency, the first principal component acquired by principal component analysis (PCA) is employed to extract the Gabor textures. In addition, for the sample patches in the BOVW model, we calculate the mean and standard deviation of each spectral/Gabor band as the spectral/textural feature. In the same way, we also generate the spectral and textural feature vectors for direct classification with SVM, as a benchmark.

Experimental Data and Setup

Datasets

The experiments were carried out on four datasets, including one remotely sensed image from the WorldView-2 (WV-2) satellite and three images from Google Earth®. The details of the four datasets are listed in Table 1, and the RGB images are shown in Figure 6 (a). Dataset 1 was acquired from WV-2, having eight spectral bands of a 2-m spatial resolution. Dataset 1 and dataset 2 were obtained at Longyan, Fujian province, where the tea that is cultivated is mainly Oolong tea. The tea in dataset 3 is Longjing tea, which is a famous green tea mainly planted in Hangzhou, Zhejiang province. The area of dataset 4 is Puer, Yunnan province, which is famous for the cultivation of Puer tea.

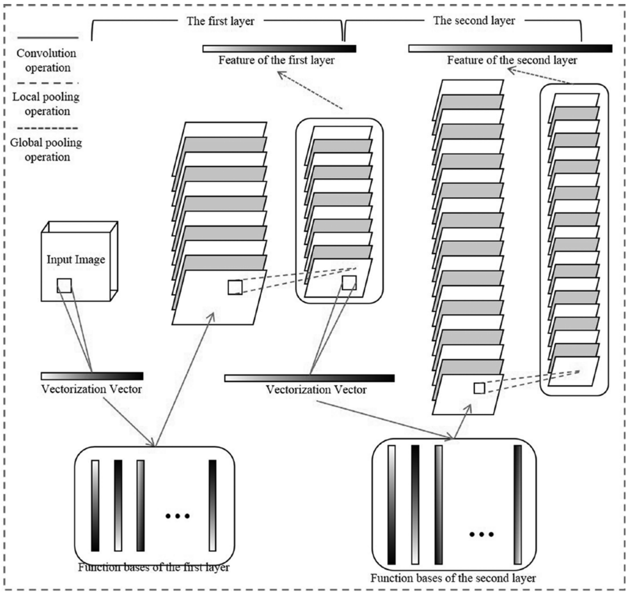


Figure 4. Framework of unsupervised feature learning.

Experimental Setup

The aforementioned datasets with different varieties of tea were used to test the performance of the proposed scene-based framework. For each testing image, 10 sets of samples were randomly chosen, with each one containing 20 tea garden scenes and 20 non-tea garden scenes as training data to train the linear SVM and sLDA, and 200 tea garden scenes and 400 non-tea garden scenes (independent of the training scenes) as test data to assess the classification accuracy. In the experiments, we selected the scenes with more than 90% of the area occupied by tea garden as tea garden scenes, and the scenes with no tea garden as non-tea garden scenes according to Huang *et al.* (2015). The scenes with less than 90% of the tea garden area were not considered in accuracy assessment. In addition, Kappa was used as the evaluation criteria, considering the unbalanced numbers of tea garden and non-tea garden scenes in the test data.

The parameters used in the four datasets are provided below:

1. Scene size: the size of each scene was 60 m × 60 m, i.e. 30 × 30 pixels for WV-2 imagery and 120 × 120 pixels for Google Earth imagery. The sensitivity of this parameter is discussed in the next section.
2. Spectral bands: 8 multispectral bands for WV-2 and 3 bands (red, green, and blue) for Google Earth images were employed to generate spectral and textural features.
3. Gabor filter: The spatial frequencies u,v are expressed in polar coordinates with radial frequency f and orientation θ. The parameters of the Gabor filter were set to $a = b = 4$, $f = \{0.006, 0.02, 0.06\}$, and $\theta = \{0, \pi/3, 2\pi/3,\}$.
4. Topic model: For the BOVW model, the size of the learned dictionary in each dataset (i.e., the number of cluster centers of k-means) was set to 100. The parameters of sLDA were determined using 5-fold cross-validation.
5. Unsupervised feature learning: In our implementation, the UCNN trained by each dataset contained two layers, with $K_1 = 100$ and $K_2 = 300$, where K_1,K_2 are the numbers of

Table 1. Details of the datasets.

	Study area	Tea	Data source	Resolution	Size	Time
Dataset1	Longyan	Oolong	WorldView-2	2m	4096×4096	2011/12
Dataset2	Longyan	Oolong	Google Earth	0.5m	4000×4000	2014/12
Dataset3	Hangzhou	Longjing	Google Earth	0.5m	4000×4000	2016/07
Dataset4	Puer	Puer	Google Earth	0.5m	4000×4000	2016/02

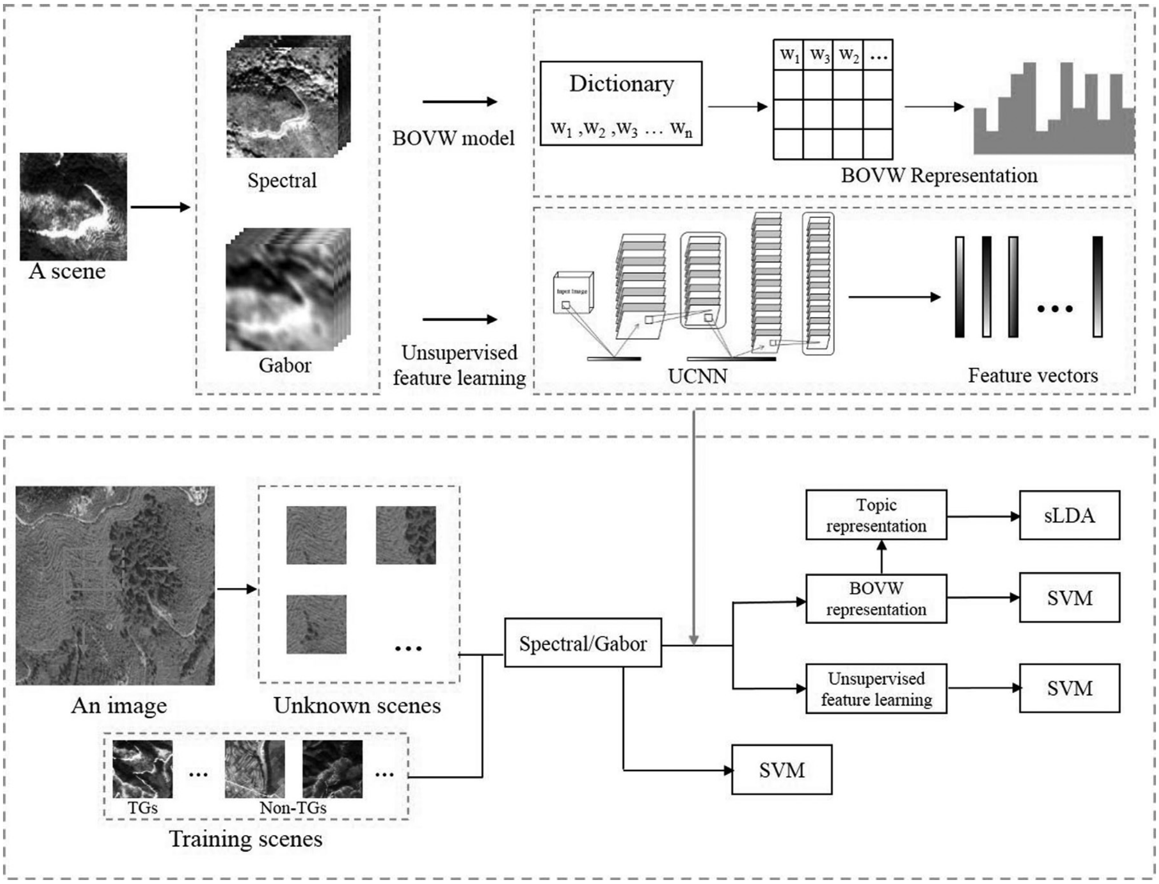


Figure 5. The proposed scene-based tea garden detection framework.

function bases in the first and second layer, respectively. The receptive field size *w* was set to 2, and the window size *s* of the local pooling operation was set to 2. All parameters were determined according to the suggestions in Li *et al.* (2016).

6. SVM: The plain linear SVM classifier, with the parameters determined by 5-fold cross-validation, was utilized for the scene classification. The training data comprised 20 tea garden scenes and 20 non-tea garden scenes.

Results and Discussion

In this section, the accuracy assessment and visual results are reported first. We then discuss the efficacy of the textual feature, and compare the different scene classification models. Finally, we analyze the sensitivity of the scene size.

General Results

In this experiment, the classification was repeated using the 10 sets of samples for all the methods and the average Kappa as well as its standard deviation was recorded, to assess the performance. The results are presented in Table 2. It can be seen that most of the Kappa values are higher than 0.80, except for Original, BOVW, and sLDA in dataset 2. Moreover, the optimal Kappa values of each dataset exceed 0.88 and are all obtained by the UCNN. These results confirm the satisfactory performance of the proposed scene-based framework for tea garden detection, and the unsupervised feature learning based UCNN performs the best of all.

Table 2. Classification accuracy (Original = SVM classification directly using the spectral or spectral-textural features).

		Dataset1	Dataset2	Dataset3	Dataset4
Spectral	Original	0.84±0.04	0.75±0.06	0.88±0.04	0.95±0.02
	BOVW	0.81±0.06	0.74±0.04	0.90±0.03	0.89±0.04
	sLDA	0.85±0.02	0.74±0.04	0.91±0.03	0.90±0.02
	UCNN	**0.88**±0.04	**0.91**±0.04	**0.94**±0.03	**0.98**±0.01
Spectral-Textural	Original	0.85±0.03	0.75±0.06	0.88±0.05	0.95±0.02
	BOVW	0.83±0.03	0.78±0.03	0.92±0.03	0.95±0.02
	sLDA	0.86±0.02	0.79±0.04	0.93±0.03	0.95±0.01
	UCNN	**0.88**±0.03	**0.88**±0.05	**0.96**±0.02	**0.98**±0.01

In order to show the detection results for the different datasets visually, we present the classification maps in Figure 6 as well as a set of examples in Figure 7. The ground-truth map for each dataset was manually delineated based on visual interpretation. In the classification maps, the scale of the minimum processing unit is half of the scene size (30 m), due to the overlapping, and in this way each unit can be covered by four scenes. The brightness of each unit is dependent on the number of tea garden scenes in the unit, i.e. the possibility of the unit belonging to tea garden (the lighter the unit, the higher the possibility that it belongs to tea garden). As shown in Figure 6, the results for dataset 2, dataset 3, and dataset 4 are very close to the ground truth, especially the maps derived from the UCNN. On the maps of dataset 1, there is a higher false alarm rate. The main reason for this is that the lower resolution of dataset 1 results in confusion between tea gardens and sparse vegetation such as orchards and bushes.

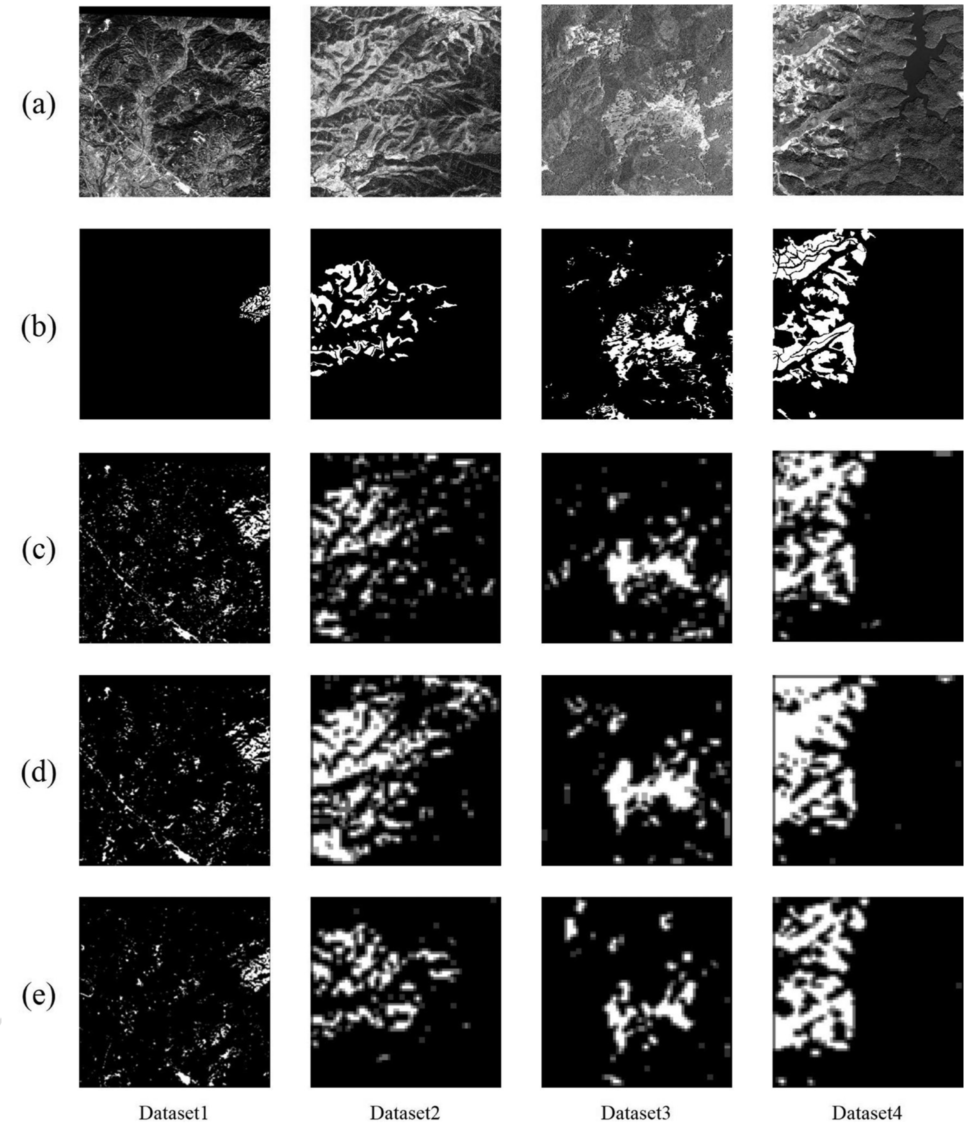


Figure 6. The classification maps: (a) Raw image, (b) Ground truth, (c) BOVW (spectral-textural), (d) sLDA (spectral-textural), and (e) UCNN (spectral).

Efficacy of Textural Features

The Kappa values of all the scene classification methods with spectral or spectral-textural features are compared in Figure 8. In general, it can be seen that the spectral-textural feature achieves comparable or better results than the spectral feature alone. For the BOVW and sLDA models, the spectral-textural feature performs significantly better than the spectral feature, and the Kappa value increases by up to 0.06. However, in terms of the UCNN, the addition of the textural feature does not raise the classification accuracy, and the Kappa value is even slightly decreased in dataset 2. The addition of the textural feature can improve the performance of the topic models, but it cannot enhance the UCNN. This phenomenon shows that the topic scene models are dependent on the input features, but the unsupervised feature learning can adaptively produce suitable and discriminative features from the original data. There is therefore no need to include manually designed features.

The visual results are shown in Figure 9, taking dataset 2 as an example. On the maps generated by BOVW, the consideration of additional textural features effectively reduces the false alarms when comparing Figure 9 (e) and (f). Figure 9 (d) shows an example image of a false alarm (bare ground, not a tea garden) in the result map of BOVW. It is incorrectly identified as a target using the spectral information alone, but is correctly classified by the textural feature. On the other hand, however, as for the UCNN model, the maps of the spectral feature and spectral-textural feature are similar, and the addition of the textural information does not change the result. As described in Li *et al.* (2016), the first and second layer in the UCNN can automatically extract the structure features such as edges, corners, and junctions, which can be regarded as the composition of texture. It is shown that these features extracted from the image by the UCNN are adequately discriminative for tea garden identification and, hence, textural features cannot increase the accuracy of the UCNN.

Comparison of Different Scene Classification Models

Figure 10 presents the best results (with spectral or spectral-textural features) for each scene classification model in the different datasets. UCNN achieves the best performance in all the datasets, especially in dataset 2, which contains a large amount of sparse vegetation. This phenomenon can also be supported by Figure 11, which shows the best visual result of each model in dataset 2. False alarms in the map of BOVW and sLDA can be observed, especially for sLDA. The classification map of the UCNN is the closest to the reference. Although the topic models have been proven to be very effective for scene classification of remotely sensed images (Lienou *et al.*, 2010; Huang *et al.*, 2015), they only achieve similar or slightly better performances than SVM classification in tea garden detection. A possible reason for this is that the handcrafted features which topic models depend on are not sufficient to represent the semantic information (i.e., the spatial relationship of sub-objects in the tea garden scenes). Instead, the UCNN, a data-driven feature descriptor, can robustly characterize the structure information and the spatial pattern of the tea plantation, which determines the semantic category of tea garden. Therefore, unsupervised feature learning based on the UCNN outperforms the topic models for tea garden detection.

Effect of the Scene Size

In the above experiments, 60 m was empirically chosen as the scene size, and the accuracies, as well as the visual results, can be regarded as satisfactory. However, the scene size can have a significant effect on the results of the proposed scene-based approach. A small scene size can preserve more details, but it is insufficient to characterize the spatial pattern of the tea gardens. On the other hand, as exhibited in Figure 12, the map with a large scene size (80 m) shows a better accuracy than the one with a small scene size (40 m), but it loses more information. In order to investigate the effect of scene size in terms of both accuracy and detail preservation, additional experiments were conducted with a series of scene sizes: 40 m, 50 m, 60 m, 70 m, and 80 m. More specifically, the best Kappa value was used to represent the detection accuracy, and mutual information entropy (Susaki *et al.*, 2014; Huang *et al.*, 2017), which measures the dissimilarity of the entropy between the two images, was employed to evaluate the detail loss. Since the map with the smallest scene size (40 m) lost the least details, it was selected as the benchmark to calculate the mutual information entropy. Figure 13 shows the results of BOVW and the UCNN in dataset2. As expected, the larger scene size results in a higher Kappa value, but a lower mutual information entropy (loss of details). It can be seen that the Kappa value increases slowly as the scene size increases, but the mutual information

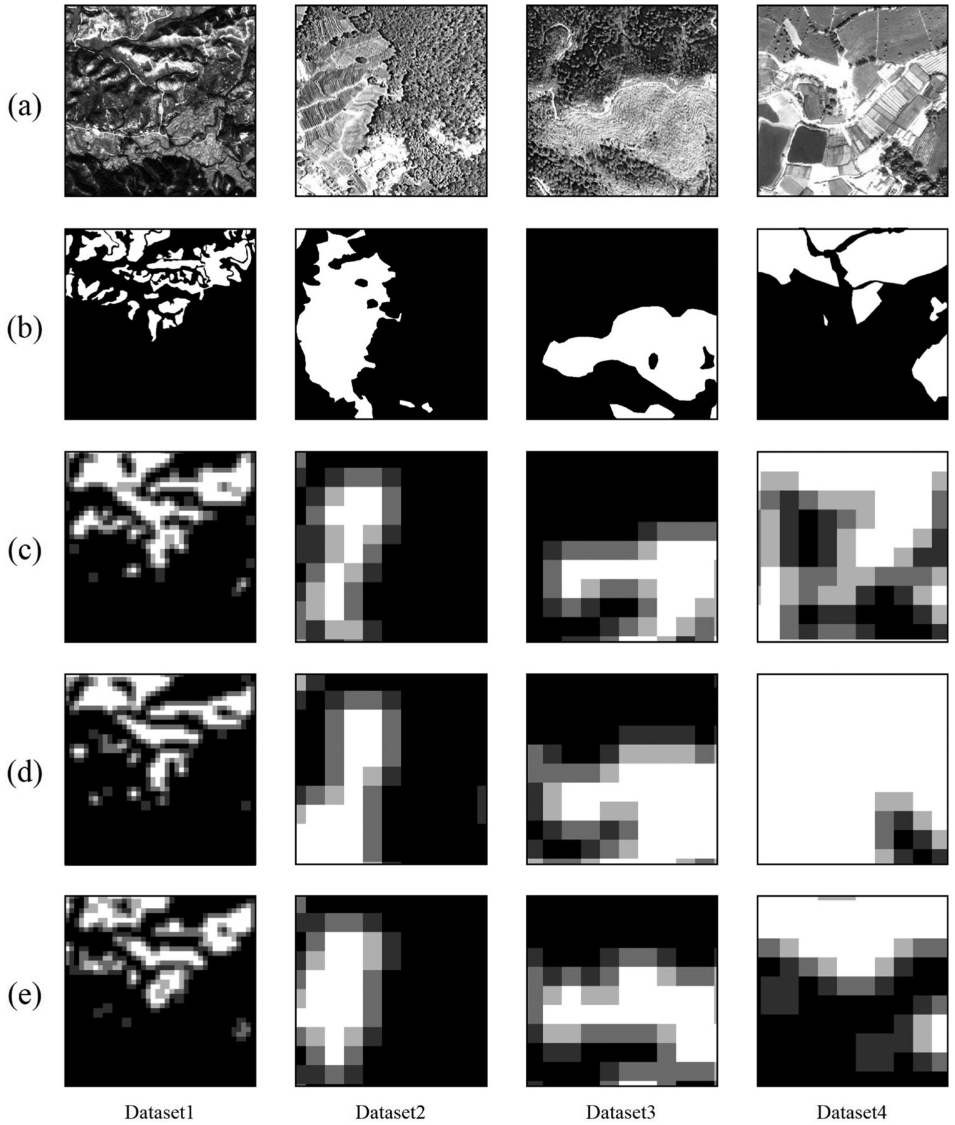
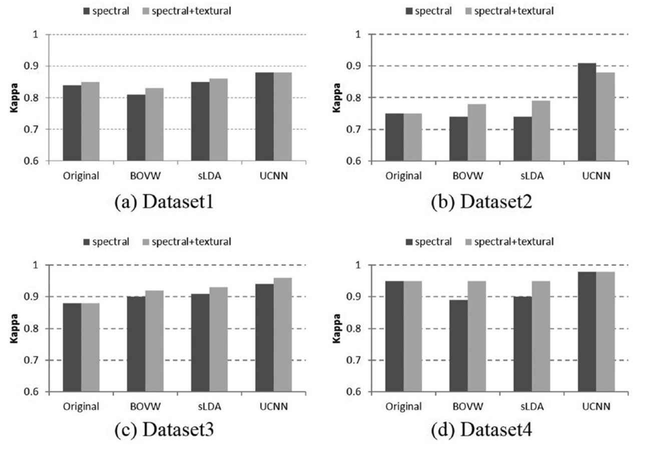
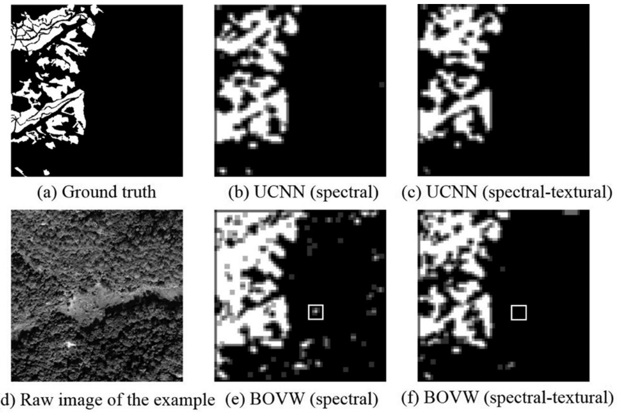


Figure 7. The example patches of the classification maps: (a) Raw image, (b) Ground truth, (c) BOVW (spectral-textural), (d) sLDA (spectral-textural), and (e) UCNN (spectral).



(a) Dataset1 (b) Dataset2 (c) Dataset3 (d) Dataset4

Figure 8. Comparison between the Kappa values of the different features.



(a) Ground truth (b) UCNN (spectral) (c) UCNN (spectral-textural)

(d) Raw image of the example (e) BOVW (spectral) (f) BOVW (spectral-textural)

Figure 9. Comparison between the classification maps of the different features in dataset 4, (d) is the raw image of the example marked by a rectangle in (e) and (f).

entropy sharply increases from 60 m to 70 m. Therefore, in this study, we selected 60 m as an appropriate scene size to balance the detection accuracy and mapping detail.

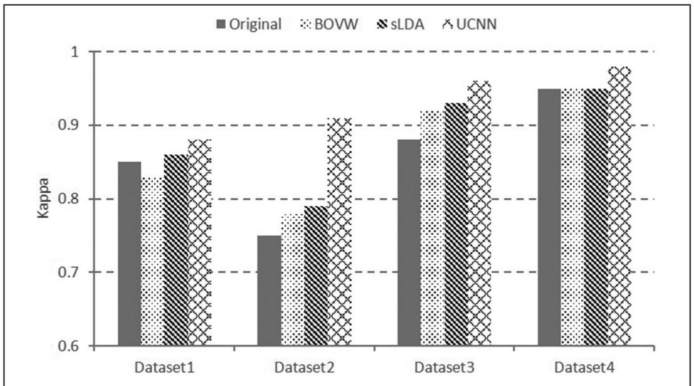


Figure 10. Comparison between the Kappa values of the different models.

Conclusions

In this paper, we have proposed a scene-based framework to effectively detect tea gardens from high-resolution remotely sensed imagery. The proposed framework is made up of three scene classification models: the bag-of-visual-words model, supervised latent Dirichlet allocation, and the unsupervised convolutional neural network. These models achieved a high accuracy and produced fine classification maps in our study area. Through discussion and comparison, it was found that unsupervised feature learning based on the UCNN outperformed the other models in both accuracy assessment and visual results. Furthermore, it was found that the supplement of the textural features could significantly improve the performance of the topic models, but it had no effect on the performance of the UCNN, since this model can automatically and adequately extract the discriminative features from the raw images.

To the best of our knowledge, this is the first study of tea garden detection using remotely sensed imagery. Since tea plants are the main cash crop in Chinese rural areas, particularly in the mountain villages in the southern provinces, tea garden detection and monitoring is important for the assessment and guidance of economic development in these areas. In the future, we plan to implement the proposed tea garden detection framework at a larger scale for tea cultivation and economic assessment.

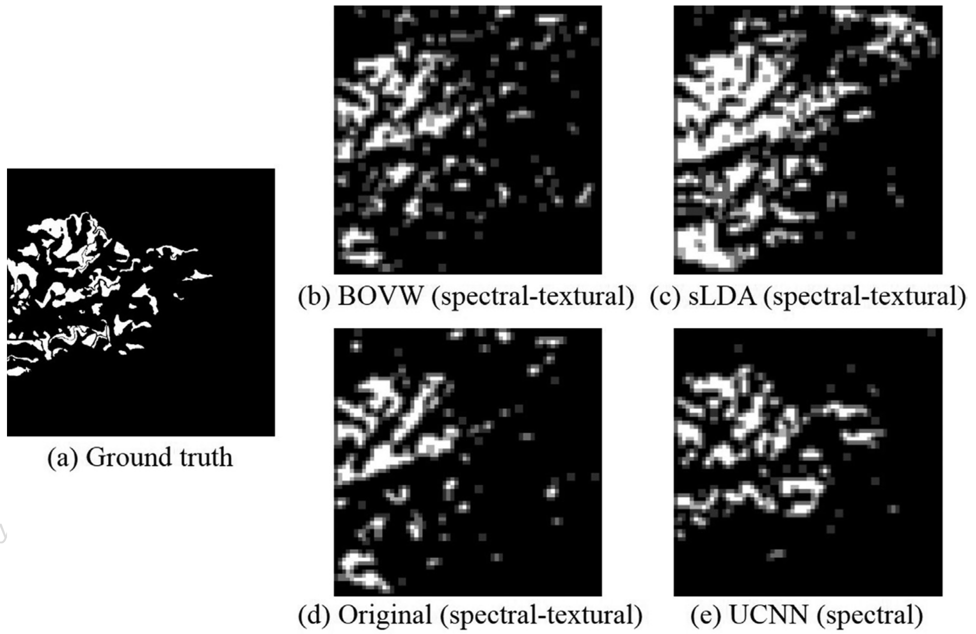

(a) Ground truth
(b) BOVW (spectral-textural)
(c) sLDA (spectral-textural)
(d) Original (spectral-textural)
(e) UCNN (spectral)

Figure 11. Comparison between the classification maps of the different models in dataset 2.

Acknowledgments

The research was supported by the National Natural Science Foundation of China under Grants 41522110 and 41771360, the Hubei Provincial Natural Science Foundation of China under Grant 2017CFA029, and the National Key Research and Development Program of China under Grant 2016YFB0501403.

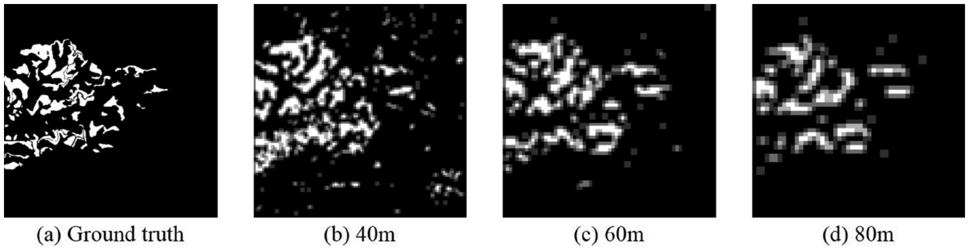

(a) Ground truth
(b) 40m
(c) 60m
(d) 80m

Figure 12. Comparison between the classification maps of differentscene sizes in dataset 2.

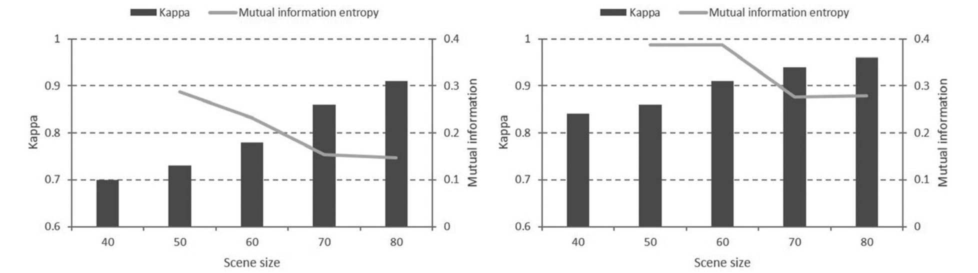


Figure 13. Comparison of the Kappa values and mutual information entropy between different scene sizes in dataset 2.

References

Blei, D.M., A.Y. Ng, and M.I. Jordan, 2003. Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3(Jan):993–1022.

Blum, M., S. Jost Tobias, J. Wülfing, and M. Riedmiller, 2012. A learned feature descriptor for object recognition in RGB-D data, *Proceedings of the 2012 IEEE International Conference on Robotics and Automation*, pp. 1298–1303.

Cabrera, C., R. Artacho, and R. Giménez, 2006. Beneficial effects of green tea-A review, *Journal of the American College of Nutrition*, 25(2):79–99.

Cheriyadat, A.M., 2014. Unsupervised feature learning for aerial scene classification, *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):439–451.

Chong, W., D. Blei, and F.F. Li, 2009. Simultaneous image classification and annotation, *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 20-25 June 2009, pp. 1903–1910.

Coates, A., H. Lee, and A.Y. Ng, 2010. An analysis of single-layer networks in unsupervised feature learning, *Ann Arbor*, 1001(48109):2.

Damodaran, B.B., J. Höhle, and S. Lefèvre, 2017. Attribute profiles on derived features for urban land cover classification, *Photogrammetric Engineering & Remote Sensing*, 83(3):183–193.

Dosovitskiy, A., J.T. Springenberg, M. Riedmiller, and T. Brox, 2014. Discriminative unsupervised feature learning with convolutional neural networks, *Advances in Neural Information Processing Systems*, 2014:766–774.

Dutta, R., A. Stein, E.M.A. Smaling, R.M. Bhagat, and M. Hazarika, 2010. Effects of plant age and environmental and management factors on tea yield in northeast India, *Agronomy Journal*, 102(4):1290–1301.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning*, 42(1):177–196.

Huang, X., and L. Zhang, 2013. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):257–272.

Huang, X., Q. Lu, and L. Zhang, 2014. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas, *ISPRS Journal of Photogrammetry and Remote Sensing*, 90:36–48.

Huang, X., H. Liu, and L. Zhang, 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 53(7):3639–3657.

Huang, X., D. Wen, J. Li, and R. Qin, 2017. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery, *Remote Sensing of Environment*, 196:56–75.

Jon, D.M., and M.B. David, 2008. Supervised topic models, *Advances in Neural Information Processing Systems*, 121–128.

Lee, T.S., 1996. Image representation using 2D Gabor wavelets, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971.

Li, Y., C. Tao, Y. Tan, K. Shang, and J. Tian, 2016. Unsupervised multilayer feature learning for satellite image scene classification, *IEEE Geoscience and Remote Sensing Letters*, 13(2):157–161.

Lienou, M., H. Maitre, and M. Datcu, 2010. Semantic annotation of satellite images using latent Dirichlet allocation, *IEEE Geoscience and Remote Sensing Letters*, 7(1):28–32.

Ming, D., X. Zhang, M. Wang, and W. Zhou, 2016. Cropland extraction based on OBIA and adaptive scale pre-estimation, *Photogrammetric Engineering & Remote Sensing*, 82(8):635–644.

Pesaresi, M., 2008. Textural analysis of coca plantations using remotely sensed data with resolution of 1 metre, *International Journal of Remote Sensing*, 29(23):6985–7002.

Putthividhy, D., H.T. Attias, and S.S. Nagarajan, 2010. Topic regression multi-modal Latent Dirichlet Allocation for image annotation, *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 13-18 June 2010, pp. 3408–3415.

Qin, R., 2015. A mean shift vector-based shape feature for classification of high spatial resolution remotely sensed imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(5):1974–1985.

Qin, Y., X. Xiao, J. Dong, Y. Zhou, Z. Zhu, G. Zhang, G. Du, C. Jin, W. Kou, J. Wang, and X. Li, 2015. Mapping paddy rice planting area in cold temperate climate region through analysis of time series Landsat 8 (OLI), Landsat 7 (ETM+) and MODIS imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:220–233.

Reis, S., and K. Taădemir, 2011. Identification of hazelnut fields using spectral and Gabor textural features, *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(5):652–661.

Sheng, G., W. Yang, T. Xu, and H. Sun, 2012. High-resolution satellite scene classification using a sparse coding based multiple feature combination, *International Journal of Remote Sensing*, 33(8):2395–2412.

Sivic, J., and A. Zisserman, 2003. Video Google: A text retrieval approach to object matching in videos, *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 13–16 October 2003, pp. 1470–1477, Vol. 1472.

Susaki, J., M. Kajimoto, and M. Kishimoto, 2014. Urban density mapping of global megacities from polarimetric SAR images, *Remote Sensing of Environment*, 155:334–348.

Wang, L., S. Hao, Q. Wang, and Y. Wang, 2014. Semi-supervised classification for hyperspectral imagery based on spatial-spectral Label Propagation, *ISPRS Journal of Photogrammetry and Remote Sensing*, 97:123–137.

Vieira, M.A., A.R. Formaggio, C.D. Rennó, C. Atzberger, D.A. Aguiar, and M.P. Mello, 2012. Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas, *Remote Sensing of Environment*, 123:553–562.

Zhu, J., Y. Su, Q. Guo, and T.C. Harmon, 2017. Unsupervised object-based differencing for land-cover change detection, *Photogrammetric Engineering & Remote Sensing*, 83(3):225–236.