HIGH-RESOLUTION REMOTE SENSING IMAGE SCENE UNDERSTANDING: A REVIEW

Qiqi Zhu, Xiongli Sun, Yanfei Zhong, and Liangpei Zhang,*

State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, P. R. China
E-mail: zhuqq@cug.edu.com
*Corresponding author E-mail: zhongyanfei@whu.edu.cn

ABSTRACT

High-resolution remote sensing (HRS) image analysis is a fundamental but challenging problem. To bridge the semantic gap, scene understanding has been proposed to achieve higher-level interpretation, through classifying the HRS scene through spatial relationship cognition and semantic induction between the land-cover objects. As a new research field, however, there has not yet been a study expatiating and summarizing the current situation of scene understanding. This paper first defines the concept of scene understanding for HRS imagery, which is different from natural image scene classification. The theory of scene understanding for HRS imagery is investigated, and is classified into four main categories: 1) scene classification based on semantic objects; 2) scene classification based on mid-level features; 3) scene classification based on deep learning; and 4) scene understanding applications based on geographic data mining.

***Index Terms*— Scene understanding, remote sensing, high spatial resolution, deep learning, semantic objects**

1. INTRODUCTION

With the appearance and rapid development of special satellite sensors, huge quantities of high-resolution remote sensing (HRS) images are now available. This type of data demonstrates the phenomenon of a complex spatial arrangement with high intra-class and low inter-class variability, which poses a big challenge for image classification. Object-based and contextual-based classification methods combining spectral-spatial features have been proposed for HRS imagery [1]. However, in many real applications, the classification of an image scene may be of more interest to the user. Commercial and residential "scenes" consist of the same kinds of land-cover classes, i.e., vegetation, buildings, and roads. However, the spatial distributions and semantic spatial relationships between the objects of these two scenes differ. It is a challenging task to recognize the scene categories using the object-based classification methods, as a result of the existence of the so-called "semantic gap".

In order to bridge the semantic gap and acquire the semantic scene information in accordance with human cognition, scene understanding has been proposed. Scene understanding relies on the cognitive understanding of the spatial relations between the different land-cover objects, and is aimed at automatically labeling an image from a set of semantic categories [2], to achieve a higher level of image interpretation. To date, scene understanding has been systematically studied in natural image analysis [3]. For HRS image analysis, scene understanding is a challenging task owing to the ambiguity and variability of the scenes.

The main objective of this paper is to present the recent advances in the techniques for the scene understanding of HRS images. In this paper, we address the following issues: (1) Concept. The concept of scene understanding for HRS imagery differs from that of natural image scene classification. (2)Theory. The theoretical basis of scene understanding for HRS imagery is investigated, including scene classification based on semantic objects, scene classification based on mid-level features, and scene classification based on deep learning. (3) Applications. Scene understanding applications based on geographic data mining are described, and we focus on urban functional zone analysis in the field of urban remote sensing. (4) The experimental results obtained using the public scene dataset are investigated, showing the differences between the different scene understanding methods.

2. HIGH-RESOLUTION REMOTE SENSING IMAGE SCENE UNDERSTANDING METHODS

Scene understanding is a rich area, consisting of segmentation, pixel labeling, object localization, scene classification, and understanding of the remote sensing images from a local to a global perspective. In this section, the HRS image scene understanding methods are reviewed. Scene understanding can be divided into four main scenarios: 1) scene classification based on semantic objects; 2) scene classification based on mid-level features; 3) scene classification based on deep learning; and 4) scene understanding applications based on geographic data mining.

2.1. Scene Classification Based on Semantic Objects

By using prior knowledge, scene classification based on semantic objects can explain the interior composition of the scene. Currently, the classification methods integrating spatial information are widely used in HRS image classification. These methods include spectrum transform [4], conditional random field [5], and object orientation based methods. Object-oriented classification methods are more suitable to construct the relations of objects for it can have clear objects. Object-oriented classification methods include segmentation and classification. After acquiring objects, many methods are designed to mine visual context relations and semantic context relations among objects [6]. Visual context is the low-level association, including spectral, texture and geometrical. Spatial context is widely used semantic context, representing high-level association, including co-occurrence relations, meaning the categories of objects being relevant and position relations, meaning the regular distribution of the objects. Relations between objects are used as features to recognize scenes. Therefore, common classifiers such as SVM, Bayesian network, and random forest are often used for scene classification.

2.2. Scene Classification Based on Mid-Level Features

The scene classification methods based on semantic objects are able to describe scenes by constructing the spatial topological relationships. However, these methods demand prior knowledge, and the final results are influenced by the accuracies of the land-cover classification. Due to the diversity of objects and the complex spatial distribution in HRS images, scene classification methods based on mid-level features have been proposed. By extracting the local features of the scenes, scene classification based on mid-level features maps the local low-level features to the corresponding parameter space to obtain the mid-level features, which have a stronger descriptive power. The methods based on mid-level features have been mainly developed from the BoVW model, which discards the spatial relationship of the objects. Feature coding methods have been developed, but these methods result in a new problem in that the dimensions of the features are very high. Hence, methods based on the PTM have been proposed.

The PTM transfers the BoVW representation of the image to a small number of topics, which are learned from the mid-level features of the scene through probabilistic statistical theory. Differing from the BoVW model, the mid-level semantic features captured from the scenes by the PTM conform more to the scene understanding of humans. Hence, the PTM, which represents the image with fewer and more effective topic features, has been receiving more and more attention. The PTMs, including the classical probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) models, introduce a latent variable to analyze the visual words generated by the BoVW model. The PTM was originally proposed as an algorithm for discovering the main topics that pervade a large and otherwise unstructured collection of documents [7]. In recent years, more and more researchers have employed PTMs to solve the challenges of HRS image scene classification.

2.3. Scene Classification Based on Deep Learning

Deep learning is an automatic feature learning and representation framework which extracts the features from the HRS imagery scenes in a joint spatial-spectral manner [8], [9]. The strategies of HRS imagery scene understanding based on a deep learning model can be divided as follows :

1) Training from scratch: training CNN on HRS imagery data with networks weights random initialization [10].

2) Fine-tuning: a pre-trained CNN net (using natural image datasets for pre-training) is used, adapting only a certain number of high-level layers [11].

3) Feature vector: take pre-trained CNN net (using natural image datasets for pre-training) as a feature extractor [12].

2.4. Scene Understanding Applications Based on Geographic Data Mining

Scene understanding results based on only HRS data may not be enough to put into practical use, and the ambiguous, broken, and irregular boundaries interfere with the identification of functional zones. The method of overlaying road vector data can solve the mosaicing problem in scene understanding. This procedure can be implemented as a pretreatment or a post-processing of the scene classification. Given an HRS image, it can be segmented into several mixed or pure scenes according to the road network, where the scenes in each block are classified with a common scene classification method [13]. The segmented scenes can then be merged to acquire the complete urban functional zone division. The accuracy of the scene classification is largely determined by the integrity of the road vector data. At present, road vector data for cities can be obtained from OpenStreetMap (OSM) or automatically extracted from the high-resolution imagery. In addition, some other social media data, such as point of interest (POI) data and real-time Tencent user density (RTUD) data, can be used to complement the extracted features and provide additional information for land-use interpretation.

3. EXPERIMENTS AND ANALYSES

3.1. Dataset description

The Wuhan IKONOS dataset was used for the scene annotation experiment. The images in the Wuhan IKONOS dataset were acquired over the city of Wuhan in China by the IKONOS sensor in June 2009. The spatial resolutions of the panchromatic images and the multispectral images are 1 m and 4 m, respectively. All the images in the Wuhan IKONOS dataset were obtained by Gram-Schmidt pan-sharpening with ENVI 4.7 software. In the Wuhan IKONOS dataset, eight scene classes are defined, namely, dense

residential, idle, industrial, medium residential, parking lot, commercial, vegetation, and water (Fig. 1). Each class contains 30 training images with a size of 150 × 150 pixels and a 1-m spatial resolution. In addition, a large image with a size of 8250 × 6150 pixels was used for the annotation (Fig. 2).

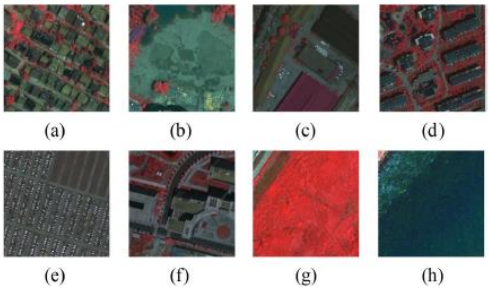


Fig. 1. Training images of the Wuhan IKONOS dataset. (a)–(h): dense residential, idle, industrial, medium residential, parking lot, commercial, vegetation, and water.

Fig. 2. Large image in the Wuhan IKONOS dataset for annotation.

3.2. Experiment settings

For the Wuhan IKONOS dataset, the large image was split into a set of small overlapping images of 150×150 pixels, with a spatial resolution of 1 m. The classification experiment produced good results when the overlap between two adjacent small images was set to 50 pixels. The different methods were evaluated using the evaluation method published in [14], where 80% of the labeled images in the Wuhan IKONOS dataset were used as training images, and the remaining images were used for testing to evaluate the model. After classification, the scene categories of the small images were used to label the large image. The final labels of the overlapping regions were decided according to the majority voting method. The different methods were executed 100 times by random selection of training samples for the three datasets. The results are reported in terms of the average value and standard deviation of the 100 runs.

In the experiments based on semantic objects, scene understanding framework based on the multi-object spatial context relationship model (MOSCRF) is applied to compare the performance with other methods. For the experiments based on mid-level features, 80% of image samples were selected per class as training samples from the Wuhan IKONOS dataset and the remaining samples were retained for testing. The PTM of LDA [14], which employs the mean and standard deviation (MeanStd) of the spectral value as the local feature descriptor, was applied to compare the performance of the different methods. The number of topics in LDA for Wuhan IKONOS dataset is optimally set as 210. In these experiments, SVM with histogram intersection kernel (HIK) was used as the classifier. For the experiments based on deep learning, the number of training samples was 80% for the Wuhan IKONOS datasets. For the CNN architecture, one of the most commonly used networks GoogleLeNet with the feature vector strategy are chosen.

3.3. Results and analysis

Fig. 3. represents scene annotation results for Wuhan IKONOS dataset based MOSCRF [7], topic model [6], and GoogleLeNet [12]. Compared with the ground truth, the methods based on mid-feature and deep learning achieve good performance, while the method based on semantic objects is not good enough for its dependency on the land cover classification results.

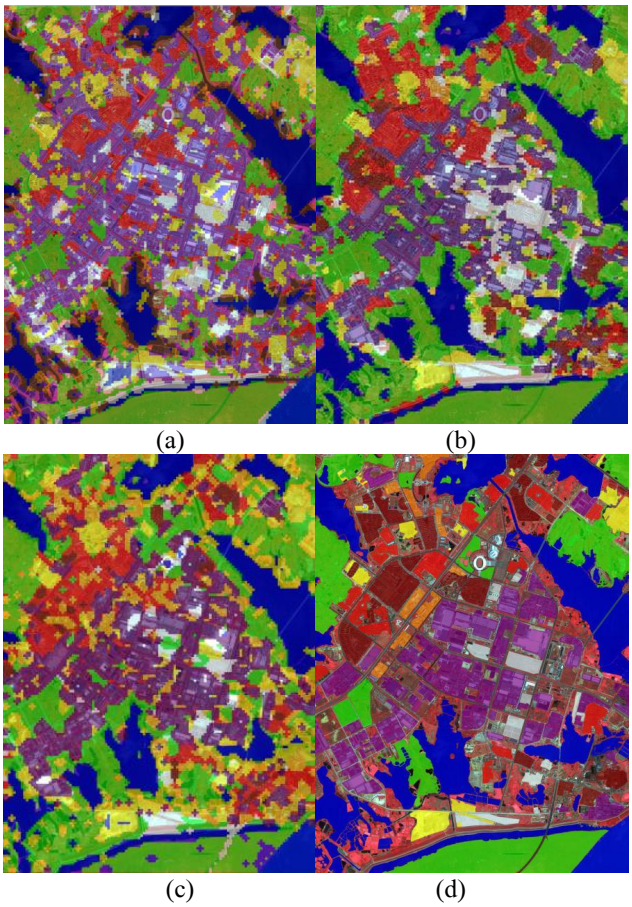
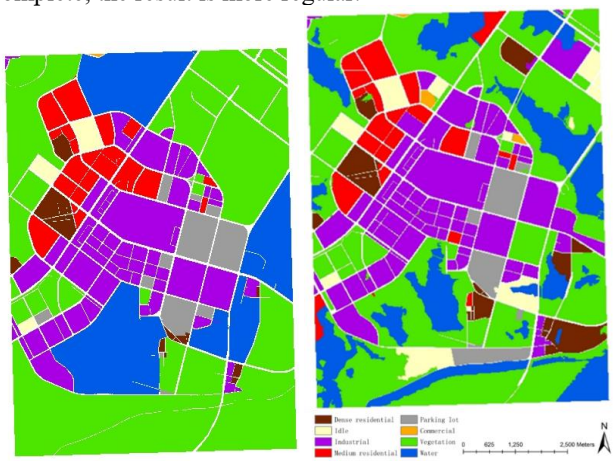


Fig. 3. Scene classification annotation results for Wuhan IKONOS dataset based on different methods. (a) MOSCRF, (b) topic model, (c) GoogleLeNet, and (d) ground truth.

When the scene unit is divided only by the road network, and the remote sensing features of each scene are extracted for classification, the result based on Fig. 3(b) is as shown in

Fig. 4(a), where the white areas represent the road network. Compared with the annotation results in Fig. 3(b), it can be seen that the mosaic phenomenon is improved, and the scene boundaries are more geographically significant. But some obvious scenes (such as water) are not classified. In order to obtain more accurate and aesthetic results, water and vegetation boundaries are extracted based specific indexes (i.e. RVI and NIR), combined with the road network constitute relatively complete scene boundaries. In addition, the traditional remote sensing features and socio-economic features obtained from crowdsourced data are taken into account when extracting scene features, which is more conducive to geographical practical application. The improved result is displayed in Fig. 4(b). Although there are some errors compared to the scene classification result, each scene is well defined when combined with the geographic data. Especially in the area where the road network is more complete, the result is more regular.



(a) (b)

Fig. 4. The scene annotation result for Wuhan combined with geographic data: (a) is the scene understanding result based on HRS and road network, (b) is the scene understanding result based on multi-source geographic data.

4. CONCLUSION

As a result of the semantic gap between the low-level features and high-level semantic concepts, it is a big challenge to obtain scene understanding results which can be directly applied to the practical use of remote sensing data. Thus, precise techniques for the modeling of scenes are of great significance for data analysis and interpretation. In this technical tutorial, we have systematically reviewed the state-of-the-art scene understanding techniques in high-resolution remote sensing (HRS) data processing. The key techniques were classified into four main categories: 1) scene understanding based on semantic objects; 2) scene understanding based on mid-level features; 3) scene understanding based on deep learning; and 4) scene understanding applications based on geographic data mining. In this review, we have surveyed the strengths and weaknesses of these methods, both theoretically and experimentally.

5. REFERENCES

[1] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, “Object bank: An object-level image representation for high-level visual recognition,” Int. J. Comput. Vis., vol. 107, no. 1, pp. 20–39, Mar. 2014.

[2] A. Bosch, X. Muñoz, and R. Martí, “Which is the best way to organize/classify images by content?,” Image Vis. Comput., vol. 25, no. 6, pp. 778–791, Jun. 2007.

[3] B. Fernando, E. Fromont, and T. Tuytelaars, “Mining mid-level features for image classification,” Int. J. Comput. Vis., vol. 108, no. 3, pp. 186–203, Jul. 2014.

[4] X. Huang, L. Zhang, and P. Li, “A multiscale feature fusion approach for classification of very high resolution satellite imagery based on wavelet transform,” International Journal of Remote Sensing, vol. 29, no. 20, pp. 5923-5941, 2008.

[5] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, “Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[6] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene Classification Based on the Sparse Homogeneous–Heterogeneous Topic Feature Model," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 5, pp. 2689-2703, 2018.

[7] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised Representation Learning for Remote Sensing Image Classification," IEEE Geosci. Remote Sens. Lett., vol. 14, no. 11, pp. 2092-2096, 2017.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Adv. Neur. In., pp. 1097–1105, 2012.

[9] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[10] F. Zhang, B. Du, and L. Zhang, “Scene classification via a gradient boosting random convolutional network framework,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[11] K. Nogueira, O. A. Penatti, and J. A. dos Santos, “Towards better exploiting convolutional neural networks for remote sensing scene classification,” Pattern Recognit., vol. 61, pp. 539–556, Jan. 2017.

[12] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, “Land use classification in remote sensing images by convolutional neural networks,” arXiv preprint arXiv:1508.00092, Aug. 2015.

[13] X. Zhang and S. Du, “A linear Dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings,” *Remote Sens. Environ.*, vol. 169, pp. 37–49, Nov. 2015.

[14] 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.