Int J Appl Earth Obs Geoinformation 70 (2018) 1–12

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

journal homepage: www.elsevier.com/locate/jag

Using multi-level fusion of local features for land-use scene classification with high spatial resolution images in urban coastal zones

Chen Lu[a,b], Xiaomei Yang[a,c,*], Zhihua Wang[a,b], Zhi Li[d]

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science, Beijing 100101, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[d] Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China

ARTICLE INFO

Keywords:
Land-use scene classification
Local features fusion
Multi-level
Urban coastal zones

ABSTRACT

Monitoring scene-level land-use in urban coastal zones has become a critical and challenging task, due to the rising risk of marine disasters and the greater number of scene classes such as harbors. Since street blocks are physical containers of different classes of land-use in urban zones, some scene classification methods based on high-spatial-resolution remote sensing images take blocks segmented by roads as classification units. However, these methods extract handcrafted low-level features from remote sensing images, limiting their ability to represent street blocks. To extract semantically meaningful representations of street blocks, the sparse auto-encoder (SAE) model was employed for local feature extraction in this paper and a multi-level method based on the fusion of local features was proposed for block-based land-use scene classification in urban coastal zones. First, convolved feature maps of street blocks were extracted by taking the hidden layer of the SAE as convolution kernels. Then, the local features were fused at three levels to generate more robust and discriminative representations of patches in convolved feature maps. The combination patterns and the absolute relationship of local features were captured at the first and second level, respectively. A convolution neural network was utilized to make the local features more discriminative to semantic information at the third level. Finally, the bag-of-visual-words model was employed to generate global features for street blocks. The proposed method was tested for Qingdao, China using Gaofen-2 (GF-2) satellite images and an overall accuracy of 83.80% was achieved in the study area. The classification results indicate that the proposed method in concert with GF-2 images has potential for accurately monitoring land-use scenes in urban coastal zones.

1. Introduction

Coastal ocean disasters occur as the result of the interaction between disaster-inducing factors (e.g. storm surges, tsunami, and sea ice), which are mainly responsible for hazard occurrence, and disaster-bearing bodies (e.g. human, property or social system), which are affected by coastal hazards. The irrational spatial distribution of disaster bearing bodies in urban coastal zones is a crucial factor in generating damage assets or threats to the safety of coastal populations (Burby, 1998; Frazier et al., 2010). Traditional methods of extracting land-cover features from remote sensing images provide us with an effective way to obtain the distribution of disaster-bearing bodies. However, the information obtained related to disaster-bearing bodies is limited. Only the category and location of a single disaster-bearing body can be known from remote sensing images. Consider a single building as an example: only the spatial location of the building can be obtained, but nothing can be known about the number of people who live in it or its economic value. With recent improvements in spatial resolution, remote sensing images can be understood more comprehensively and deeply. The scene-level land-use categories such as residential, industrial, and commercial areas can be assigned to a sub-image containing several disaster-bearing bodies, to provide us with clues about the distribution of the human population and property. As a result, a more detailed exposure assessment of disaster-bearing bodies can be performed in urban coastal zones (Taramelli et al., 2014; Zhang et al., 2016; Rizzi et al., 2017).

This creates an important demand for a method that allows us to automatically extract scene-level urban land-use from high-spatial-

* Corresponding author at: State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science, 11A Datun Road, Chaoyang District, Beijing 100101, China.
E-mail addresses: luchen@lreis.ac.cn (C. Lu), yangxm@lreis.ac.cn (X. Yang).

https://doi.org/10.1016/j.jag.2018.03.010
Received 29 September 2017; Received in revised form 25 March 2018; Accepted 30 March 2018

resolution (HSR) remote sensing images. Methods based on land-cover features and those based on scene classification have shaped the field. Coverage ratio, density (Herold et al., 2003; Hu and Wang, 2013; Lowry and Lowry, 2014), and spatial arrangement (Zhan et al., 2002; Comber et al., 2012; Vaduva et al., 2013; Walde et al., 2014; Li et al., 2016; Zhong et al., 2017) of land-cover features in pre-defined land-use units have been used to extract scene-level land-use. Other data sources such as point-of-interest data and Open Street Map (OSM; Haklay and Weber, 2008) data have also been integrated with land-cover features for extracting land-use information (Li et al., 2017b; Zhang et al., 2017). Without relying on land-cover features, scene classification directly provides a scene-level land-use class to a subset of an image. Compared with the methods based on land-cover features, scene classification can be more easily applied to automatically extract scene-level land-use information, because this eliminates some of the cost involved in extracting land-cover features in a study area. Moreover, scene classification can potentially be adapted to all well-defined land-use classes. Although unique scene-level land-use classes such as beach and harbor exist in urban coastal zones, scenes belonging to these classes can be recognized by collecting training samples and training the existing scene classification model, avoiding designing new methods for a specific land-use class. Therefore, methods based on scene classification are more suitable than those based on land-cover features for monitoring scene-level land-use in urban coastal zones with various semantic classes.

Compared with the land-cover features types (such as water, building and road) extracted by pixel-level image classification, scene-level land-use categories (such as residential, industrial, and commercial areas) are high-level semantics. Although the low-level features (such as color, texture and shape) have been successfully applied to pixel-level image classification (Bordes and Prinet, 2008; Li et al., 2015; He et al., 2017), the low-level features sometimes cannot precisely represent the high-level semantics contained in land-use units. The main challenge in scene classification is bridging the so-called semantic gap between the low-level features and the high-level semantics. Some sophisticated models for scene classification have been developed for this purpose. By deploying these models, only low-level features extracted from local regions of an image can form a compact and relatively discriminative representation for scene classification (Chen et al., 2015; Zhu et al., 2016; Zhao et al., 2016). The typical models for scene classification applied in the remote sensing community are the bag-of-visual-words (BoVW) model (Yang et al., 2007), probability latent semantic analysis (Hofmann, 2001), and latent Dirichlet allocation (LDA; Blei et al., 2003). Another way to bridge the semantic gap is to extract local features that have more semantic meanings by using unsupervised feature learning. By conducting simple operations such as max-pooling or concatenating on local features, higher-accuracy classification models can be established (Zhang et al., 2015a; Fan et al., 2017). Approaches such as sparse coding (Pati et al., 1993), sparse auto-encoder (SAE; Poultney et al., 2007), and deep belief networks (Mohamed et al., 2012) have been applied to extract more discriminative local features.

The above research studies focused on giving a semantic label to square scene images of the same size. These methods aimed to improve the classification accuracies of standard datasets or to annotate a large HSR remote sensing image with sub-images of a fixed size as annotation units. When applying these methods in a realistic urban environment, a street block (normally an area bounded by streets on all sides) belonging to a single semantic category may be divided into several parts. However, this will raise two problems: 1) if the semantic category of the street block can be obtained by conducting classification on local parts, the parts at the edge of the street block may be classified incorrectly because these parts do not contain sufficient semantic information; 2) if the semantic category is dominated by the composition of all objects in an entire street block, the correct semantic category of the street block cannot be obtained by perceiving any local parts. The first problem can be alleviated by dividing an image into larger subsets with overlapping areas. However, the second problem cannot be resolved when the classification units have a fixed size. Therefore, the use of more effective classification units needs to be explored.

Since the above problems can be avoided by directly taking street blocks as classification units, some recent work has conducted research into scene classification based on street blocks. Zhang et al. (2015b) established a scene classification model for street blocks by measuring intra-scene feature similarity and inter-scene semantic dependency. Then, a linear Dirichlet mixture model was further developed for the quantification of mixed semantics of urban scenes (Zhang and Du, 2015). However, these methods use handcrafted low-level features such as spectral histograms, or gray level coexistence matrices to extract features from HSR remote sensing images, limiting their ability to represent street blocks. Furthermore, land-cover features and building data are still indispensable for improving the representation of street blocks. Guided by this observation, the SAE model was employed for scene classification based on street blocks in the current study, and more semantically meaningful representations of street blocks were extracted. Hence scene classification based on street blocks can be accomplished with only HSR remote sensing images.

Several scene classification models based on SAE have been established (Zhang et al., 2015a; Cheng et al., 2015; Li et al., 2017a; Han et al., 2017). In their methods, the same number of local features extracted by SAE were further fused to one feature by different approaches such as max-pooling (Zhang et al., 2015a), mean, variance and standard deviation (Li et al., 2017a). Since the number of local features extracted from street blocks of different shapes and sizes vary, the number of local features to be fused should be set to different values to ensure that a scene image is represented by the same number of fused features and that traditional classifiers can be used. For bigger street blocks, more local features would be fused and many useful details would be lost, thus decreasing the discrimination ability of the fused features. Therefore, existing methods based on SAE are expected to not work well for scene classification based on street blocks.

To solve the problem that larger street blocks would lose more details by applying the existing methods based on SAE, more robust and discriminative local features should be extracted from street blocks, and all local features extracted from a block should equally contribute to a global feature. Therefore, a multi-level feature fusion framework is proposed based on SAE for scene classification of street blocks in this paper. In particular, the SAE was employed to extract local features of the street blocks from remote sensing images. Then three different methods were utilized to fuse the local features. The BoVW was additionally utilized to generate global features from the fused features. Finally, a support vector machine (SVM; Chang and Lin, 2011) with a histogram intersection kernel (HIK; Barla et al., 2003) was employed to classify the global features of all the street blocks.

2. Methods

In this section, a framework is proposed for scene-level land-use classification with street blocks used as classification units. The flowchart of the proposed method is shown in Fig. 1. In the following sections, the SAE model, local feature extraction, multi-level fusion, BoVW representation, and classification are described in detail.

2.1. Sparse auto-encoder

The SAE model is a symmetrical neural network with an input layer, an output layer, and a hidden layer connecting them. Useful features in the input data are learned by minimizing the squared reconstruction error between the input data at the input layer and its reconstruction at the output layer and imposing sparsity on the hidden layer.

SAE consists of two parts, an encoder and a decoder. In the encoder stage, the input $x^i \in R^N$ at the input layer is mapped to the activation $a \in R^K$ at the hidden layer, where K is the number of hidden units. The