ISPRS Journal of Photogrammetry and Remote Sensing 138 (2018) 281–294

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery

Yanfei Zhong, Xiaobing Han *, Liangpei Zhang

State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, PR China
Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, PR China

ARTICLE INFO

Article history:
Received 30 June 2017
Received in revised form 9 February 2018
Accepted 16 February 2018

Keywords:
Geospatial object detection
High spatial resolution (HSR) remote sensing imagery
Residual network
Pre-training mechanism
Position-sensitive balancing

ABSTRACT

Multi-class geospatial object detection from high spatial resolution (HSR) remote sensing imagery is attracting increasing attention in a wide range of object-related civil and engineering applications. However, the distribution of objects in HSR remote sensing imagery is location-variable and complicated, and how to accurately detect the objects in HSR remote sensing imagery is a critical problem. Due to the powerful feature extraction and representation capability of deep learning, the deep learning based region proposal generation and object detection integrated framework has greatly promoted the performance of multi-class geospatial object detection for HSR remote sensing imagery. However, due to the translation caused by the convolution operation in the convolutional neural network (CNN), although the performance of the classification stage is seldom influenced, the localization accuracies of the predicted bounding boxes in the detection stage are easily influenced. The dilemma between translation-invariance in the classification stage and translation-variance in the object detection stage has not been addressed for HSR remote sensing imagery, and causes position accuracy problems for multi-class geospatial object detection with region proposal generation and object detection. In order to further improve the performance of the region proposal generation and object detection integrated framework for HSR remote sensing imagery object detection, a position-sensitive balancing (PSB) framework is proposed in this paper for multi-class geospatial object detection from HSR remote sensing imagery. The proposed PSB framework takes full advantage of the fully convolutional network (FCN), on the basis of a residual network, and adopts the PSB framework to solve the dilemma between translation-invariance in the classification stage and translation-variance in the object detection stage. In addition, a pre-training mechanism is utilized to accelerate the training procedure and increase the robustness of the proposed algorithm. The proposed algorithm is validated with a publicly available 10-class object detection dataset.

1. Introduction

Machine vision is a high-tech field of electronics in which different types of applied processing are performed, in applications in the earth observation, geoscience, remote sensing, and radar and medical image processing fields (Akbarizadeh, 2012; Karimi et al., 2017a, 2017b; Modava and Akbarizadeh, 2017; Rahmani and Akbarizadeh, 2015; Tirandaz and Akbarizadeh, 2016). In recent years, the rapid development of high spatial resolution (HSR) remote sensing sensors has supplied us with abundant detail and spatial structural information, thereby providing us with new opportunities for advancing the automatic interpretation of remote sensing images (Cheriyadat et al., 2014; Cheng et al., 2013, 2014, 2016a, 2016b; Eikvil et al., 2009; Lei et al., 2012; Lin et al., 2016; Liu and Shi, 2014; Liu, 2014, Liu et al., 2011; Siva et al., 2012; Yao et al., 2017; Yu et al., 2015; Zhang et al., 2015, 2016a, 2016b, 2017; Zhao et al., 2016a). Clearly, the HSR remote sensing sensors are one of the machine vision applications in the electronics field and there are many real applications, such as land-use/land-cover (LULC) classification (Zhao et al., 2018; Zhong et al., 2014; Han et al., 2017c), object detection (Zhong and Wang, 2007; Han et al., 2014, 2015, 2017b), scene classification and understanding

* Corresponding author at: State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, PR China; Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, PR China.

E-mail address: whu_hxb@whu.edu.cn (X. Han).

https://doi.org/10.1016/j.isprsjprs.2018.02.014
0924-2716/© 2018 Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS).

(Han et al., 2017a, 2017d; Zhao et al., 2016b; Zhong et al., 2015, 2016; Cheng et al., 2017a, 2017b).

Multi-class geospatial object detection, being a fundamental but challenging problem in the field of remote sensing imagery analysis, has recently attracted considerable attention and is becoming more and more important in a wide range of civil and engineering applications (Aksoy et al., 2012; Bai et al., 2014; Blaschke et al., 2014; Blaschke, 2010; Bontemps et al., 2008; Cheng et al., 2013, 2014, 2016, 2017a, 2017b; Cheriyadat, 2014; Grabner et al., 2008; Han et al., 2015; Han et al., 2014; Li and Perona, 2005; Yang and Newsam, 2010; Yokoya and Iwasaki, 2015; Yu et al., 2015; Zhang et al., 2015; Zhong and Wang, 2007). Differing from natural imagery obtained by a camera on the ground from an horizontal view, HSR remote sensing imagery is obtained by satellite-borne or space-borne sensors from a top-down view, which is an approach that can be easily influenced by weather and illumination conditions. In addition, differing from the anteroposterior object detection from natural imagery, the position of the objects in HSR remote sensing imagery is mostly left–right (Zhao et al., 2016a). For these reasons, the core task of object detection from HSR remote sensing imagery should be clearly illustrated. Object detection is to determine whether or not a given image contains one or more objects belonging to the class of interest, and to locate the position of each predicted object in the image (Cheng et al., 2013, 2016, 2017a, 2017b; Han et al., 2014; Han et al., 2015). The object source category should also be clearly explained. Similar to the category settings of LULC classification, the object categories consist of natural modality and man-made modality, where natural modality objects usually refer to the parcels with vague boundaries that are part of the background, and man-made modality objects refer to the objects with sharp boundaries that are independent of the background environment (e.g., ships, buildings) (Cheng et al., 2014, 2016). Due to the complex ground object distribution of HSR remote sensing imagery, multi-class geospatial object detection from HSR remote sensing imagery is greatly affected by the object properties of viewpoint variation, occlusion, background clutter, illumination, shadow, and small size. The relative dearth of geospatial objects is also a big challenge for the current HSR remote sensing imagery object detection methods.

Based on the above introduction to HSR remote sensing imagery object detection, there is an urgent need to promote the performance of object detection. A number of traditional methods based on handcrafted features have been proposed for HSR remote sensing imagery object detection. The traditional object detection methods usually regard the object detection problem as a classification problem. The traditional object detection methods can be separated into three stages: the region proposal generation stage with selective search (SS) method, the feature extraction stage, and the classification stage. The SS method generates a large number of regions to detect the locations of objects. After the region proposal generation stage with SS (Uijlings et al., 2013), the manually designed features, such as the shape feature, context feature, or local image feature (e.g., scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG) (Dalal and Triggs, 2005)), are extracted at a shallow level. Finally, the extracted features are imported to a classifier such as support vector machine (SVM) (Xu et al., 2010), conditional random fields (CRF) (Zhong and Wang, 2007), a sparse-coding-based classifier (Sun et al., 2012), or a bag-of-words (BoW) classifier (Xu et al., 2010). Although the traditional object detection methods can obtain good performances, the problems are obvious. The first problem with SS is the inaccurate proposal generation, which usually generates a large number of proposals with much redundancy and consumes a lot of time. Furthermore, the human labor involvement in the feature design procedure increases the non-automatic capability, and these non-related stages reduce the efficiency of the traditional methods. To summarize, the core part of the traditional object detection methods is the feature extraction stage.

Although the traditional object detection methods can obtain relatively impressive results, the feature extraction stage is far from automatic. Deep learning is an automatic feature learning and representation framework, which can learn deep features from the data itself (LeCun et al., 1998, 2015; LeCun, 1990). Among the deep learning methods, the convolutional neural network (CNN) is an efficient and automatic hierarchical feature learning framework, which has demonstrated great potential in object detection. A series of object detection methods have been developed based on deep learning, including regional CNN (R-CNN) (Girshick et al., 2016), Fast R-CNN (Girshick, 2015), and Faster R-CNN (Ren et al., 2015). However, these object detection methods focus mainly on natural imagery object detection. R-CNN is the first deep learning based object detection framework, which realizes the CNN-based feature automatic extraction task after a region proposal generation procedure. Fast R-CNN improves object detection based on R-CNN by undertaking the region proposal generation procedure after the convolutional feature generation, to avoid the CNN operation on all the regions, which reduces the calculation time of object detection. Faster R-CNN accelerates Fast R-CNN by generating the region proposals with an automatic region proposal network (RPN), and then improves the computational efficiency by sharing the features between the region proposal generation procedure and the object detection procedure. However, in these CNN-based object detection frameworks, the recognition ability is influenced by the CNN. With the development of computational devices such as the graphics processing unit (GPU), the performance of image recognition and classification has been improved by the deeper fully convolution network (FCN) networks (LeCun et al., 2015) such as the residual network (ResNet) (He et al., 2016) and ultra-deep networks such as FractalNet (Larsson et al., 2016).

The recognition performances of the deep FCN are superior due to the specific capsulation or residual processing structures, whereas the translation with convolution operation has little influence on the categories of the HSR remote sensing scene imagery, but can greatly influence the detection performance from the HSR remote sensing imagery. However, the traditional deep learning based object detection methods cannot balance the translation-invariance between the classification stage and the object detection stage.

In order to further improve the performance of multi-class geospatial object detection from HSR remote sensing imagery, the feature sharing object detection framework has been developed for use in a more powerful feature representation model: ResNet. ResNet is a residual learning framework that can ease the training of networks that are substantially deeper than those networks used previously. For ResNet, the layers are reformulated as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. Compared with VGGNet and GoogLeNet, the representation ability of ResNet is greatly improved with the introduction of residual functions to enhance the supervision. However, due to the introduction of a region of interest (RoI) pooling layer into the deep learning based two-stage object detection framework, the balance of the translation-invariance in the classification stage and the translation-variance in the object detection stage has not been considered, which causes position accuracy problems in the detection procedure for HSR remote sensing imagery with objects of complex distribution. In order to solve the above problems in the object detection task, a position-sensitive balancing (PSB) framework based on ResNet is proposed for accurately detecting the multi-class geospatial objects in HSR remote sensing imagery. The