An Improved Bounding Box Post-processing Algorithm with Faster R-CNN for High Spatial Resolution Remote Sensing Imagery Object Detection

Yanfei Zhong[1,2], Xiaobing Han[1,2], Liangpei Zhang[1,2]*
State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing,
Wuhan University, Wuhan 430079, P.R. China
Collaborative Innovation Center of Geospatial Technology,
Wuhan University, Wuhan 430079, P.R. China
*Corresponding author E-mail: zhongyanfei@whu.edu.cn Phone: +86-27-68779969

Key words: object detection, high spatial resolution (HSR) remote sensing imagery, faster-RCNN, region proposal

Abstract: Multi-class geospatial object detection from high spatial resolution (HSR) remote sensing imagery is a profound but challenging task. Due to the powerful feature representation ability of deep learning, object detection from HSR remote sensing imagery is more and more efficient. Among the deep learning based object detection methods, region proposal based object detection method (e.g., faster region based convolutional neural network (Faster R-CNN)) is effective, which requires to generate the region proposals from the input imagery at first and then regress the locations of the bounding boxes and classify the categories. As an important and requisite procedure of Faster R-CNN, properly selecting and accurately suppressing the redundant bounding boxes is significant and critical. Traditional bounding boxes suppression operation is conducted mainly by the non-maximum suppression (NMS) operation, which only conserve the maximum score of the predicted bounding boxes and deletes the non-maximum scores. However, the NMS operation may be of little effect especially for the situations where the objects are densely distributed, the bounding box sizes are not proper, and the overlaps exist between the bounding boxes. To relieve this phenomenon in object detection fields and further improve the detection performance of the region proposal based Faster R-CNN frameworks, this paper proposed to utilize a sigmoid derivative decay function to replace the traditional NMS for Faster R-CNN. The proposed algorithm has been validated and experimented on a publicly available ten-class object detection dataset.

1. Introduction

High spatial resolution (HSR) remote sensing imagery have emerged in huge quantities nowadays, which covers a wide range of spatial resolution from tens of meters to several meters and rapidly develops towards sub meter. This property enables the abundant detail and structural information of the earth's surface to be clearly recorded by the HSR remote sensing imagery, which provides a powerful means to accurately interpret the ground objects. As a result, a wide range of applications based on HSR remote sensing imagery have been found, such as disaster control, land planning, urban monitoring, and traffic planning [1]. And one of the most promising and important prospects in HSR remote sensing imagery is geospatial object detection [2]-[3].Object detection for HSR remote sensing imagery has achieved great success in both civil and military fields, particularly in reconnaissance and surveillance applications [3].

Object detection is defined to determine whether a given aerial or satellite image contains one or more objects belonging to the class of interest and to locate the position of each predicted object in the image [3]. The term object refers to its generalized form, including natural modality and man-made modality, where the natural modality objects refer to the parcels with vague boundaries that are part of the

background and the man-made modality objects refer to the objects with sharp boundaries that are independent of the background environment (e.g., ships, buildings, athletic facilities) [3]. In a survey, enormous object detection methods based on optical remote sensing imagery have been induced and summarized into four categories: template matching-based object detection methods, knowledge-based object detection methods, object-based image analysis (OBIA)-based object detection methods, and machine learning-based object detection methods [8]. As far as our knowledge, current object detection methods for HSR remote sensing imagery usually consists of several procedures, such as feature extraction, classification, and localization, and the main stream is to convert the object detection problem into a classification problem [4]-[6]. In addition to the staged object detection pipeline, current object detection heavily rely on human-labor, not only the human-labor in labeling each bounding box to be predicted but also the handcrafted feature extractors (e.g., HOG, SIFT). To summarize, effective feature representation plays an important role for object detection.

Since the aforementioned feature representation mechanism plays a significant role in achieving high-performance object detection results for HSR remote sensing imagery, exploring an effective feature representation algorithm is an urgent task for object detection. A successful and fundamental object detection method with low-level handcrafted features is bag-of-word (BOW) model. Its essence idea is to treat each region proposals, obtained by selective search (SS) method, as a collection of unordered local descriptors, quantizes them into a set of visual words, and then computes a histogram representation. Compared with the low-level handcrafted feature representation methods, deep learning [7] offers an effective and automatic feature representation hierarchical framework for object detection, where convolutional neural network (CNN) [9] is a powerful means for high-level feature extraction. With the introduction of CNN, in natural imagery object detection field, a series of object detection works based on CNN, such as region-based convolutional neural network (R-CNN) [10] by solving the feature extraction problem with effective CNN, fast region-based convolutional neural network (fast R-CNN) by solving the repeated region proposal computation with an effective multi-scale mapping operation [11], and faster region-based convolutional neural network (faster R-CNN) [12] by solving the feature sharing between region proposal generation and fast R-CNN. Although these CNN based object detection works have achieved superior performance with automatic feature representation, properly selecting the accurate bounding boxes from a large number of predicted bounding boxes has a significant influence on the object detection performance.

As a significant and requisite procedure of region proposal based object detection methods, NMS is an important and necessary post-processing procedure. NMS operates on object detection score matrix and the coordinate information of the region by selecting the highest confidences and suppressing the non-maximum highest but significant confidences by totally deleting, including the bounding boxes with a score higher than the threshold. In addition, this procedure is recursively applied on the remaining boxes. Thus, if an object lies within the predefined overlap threshold, it leads to a miss. However, for the multi-class geospatial object detection from high spatial resolution remote sensing imagery, the object distributions are influenced by the complicated situation of the HSR remote sensing imagery such as viewpoint, complex background clutter, illumination, shadow, et al. When dealing with the densely distributed and small size objects from the high spatial resolution remote sensing imagery, neighboring windows often have correlated bounding boxes due to multi-scale design mechanism of Faster RCNN, and it may easily decay the object detection performance. In order to conserve the bounding boxes with non-maximum scores with a continuous function of their overlap and no non-maximum bounding boxes deleted, an improved bounding box post-processing Faster-RCNN object detection algorithm has been

proposed. The proposed improved Faster R-CNN framework conducts the bounding boxes selecting and suppressing by utilizing several decay functions.

In order to construct the decay function, this paper proposes to utilize the sigmoid derivative function as the decay function. Compared with the hard NMS operation, the sigmoid derivative function works by assigning a very low score when a bounding box has a very high overlap with the maximum score; otherwise, it should be assigned a low overlap when it can maintain its original detection score. By applying the improved NMS algorithm, the object detection framework can solve the situation like this. When there is a detection box which is very close to an object (within the threshold overlap), but has a slightly lower score than the maximum score (the maximum score does not cover the object), thus the detection box gets suppressed by a low threshold. Thus, when the threshold is low, the miss-rate would increase; when the threshold is high, the false positives would increase. However, the increase would be much higher than the increase in true positives because the number of objects is typically much smaller than the number of RoIs generated by the Faster R-CNN detector. The proposed improved NMS algorithm with Faster R-CNN can handle the situation like this. If the predicted bounding boxes contain an object not covered by the maximum score, it would not increase the miss-rate.

The contributions of this work can be summarized into three aspects.

1) This paper utilizes the comprehensive and powerful feature representation CNN structure, ZF and VGG, to extract the features from the high spatial resolution remote sensing imagery.
2) Considering the powerful feature extraction ability with deep learning, this paper adopts the effective Faster RCNN series methods as the baseline object detection methods. In addition, considering the effective transferring and pre-training mechanism, the Faster RCNN is applied with a powerful pre-trained Faster RCNN series methods for multi-class geospatial object detection of HSR remote sensing imagery.
3) The proposed improved I-Faster R-CNN framework combines the Faster RCNN object detection methods with a generalized and flexible bounding boxes post-processing operation. The generalized bounding boxes post-processing works mainly by constructing the decay function with a continuous function to conserve the non-maximum bounding boxes scores. Thus, the proposed algorithm can improve the object detection performance on the basis of Faster-RCNN methods for HSR remote sensing imagery.

The proposed framework was evaluated and compared with the conventional HSR remote sensing imagery object detection methods as well as the current non-end-to-end CNN based object detection methods and the Faster R-CNN series object detection methods. The experimental datasets adopted the 10-class HSR remote sensing imagery geospatial object detection dataset—NWPU VHR-10 dataset. The experimental results confirmed that the proposed method can achieve a satisfactory detection result with limited labeled training samples.

The rest of the paper is organized as follows. Sec. 2 describes the proposed algorithm. Sec. 3 evaluates the algorithm using a public 10-class object dataset. Conclusion is draw in Sec. 4.

2. The Proposed Improved Faster R-CNN Object Detection Framework for High Spatial Resolution Remote Sensing Imagery

In this section, we firstly introduce the overall architecture of the proposed improved Faster R-CNN object detection framework, then the proposed improved Faster R-CNN framework will be introduced from two aspects, namely the feature representation based on CNN, the Faster R-CNN object detection

framework with the improved bounding box post-processing algorithm with Faster R-CNN framework for HSR remote sensing imagery object detection. The overall flowchart of the proposed algorithm is shown in Figure 1.

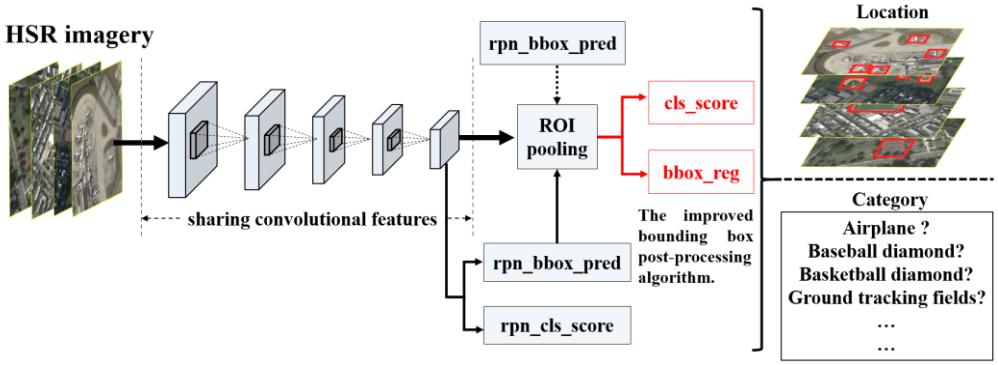

Figure 1. The overall flowchart of the proposed algorithm.

2.1 Feature extraction based on CNN

It's known that CNN is an efficient and powerful automatic feature extraction network in deep learning field, which usually consists of several convolutional layers, pooling layers, nonlinear layers et al. AlexNet, GoogLeNet, VGGNet, ResNet et al. are typical CNN architectures. Each structure has its own characteristic, but they all share similar advantages such as powerful feature extraction ability and automatic feature representation ability. For the proposed G-Faster-RCNN architecture, the Zeiler and Fergus (ZF) model and the visual geometry group (VGG) model.

2.2 Faster R-CNN with the improved bounding box post-processing algorithm

The proposed improved bounding box post-processing algorithm with Faster R-CNN can be illustrated from three stages, namely the RPN stage, the Fast R-CNN stage, and the improved bounding box post-processing algorithm.

RPN is the core innovation of the Faster R-CNN based object detection framework, which is a kind of fully convolutional network (FCN) that deals with the arbitrary-size input images and generates a set of rectangular object proposals. The characteristic of RPN is the utilization of anchors. Anchors are the centers of the centers of the sliding windows, and they construct the different-ratio and multi-scale region proposals to import into the RPN. With the anchors, the RPN can realize the multi-scale information incorporation.

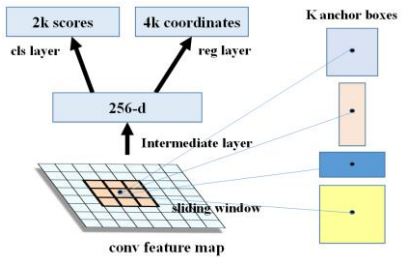

Figure 2. The principle of the RPN.

Fast R-CNN is a location refinement stage which takes an entire image as input and a set of object proposals to score. The network first processes the whole image with several convolutional and max pooling layers to produce a convolutional feature map. Then, for each object proposal, a region of interest

(RoI) pooling layer extracts a fixed-length feature vector from the feature map. Each feature vector is fed into a sequence of fully connected layers that finally branch into two sibling output layers: one is utilized to produce softmax probability estimates over K object classes plus a "background" class, and the other layer is utilized to output four real-valued number for each of the K object classes. Each set of the four values encodes refined bounding-box positions for one of the K classes.

In dealing with the HSR remote sensing imagery object detection, there are three situations should be taken into consideration: 1) The scores of neighboring detections should be decreased to an extent that they have a smaller likelihood of increasing the false positive rate, while being above obvious false positives in the ranked list of detections; 2) It's necessary to remove neighboring detections altogether with a low NMS threshold would be sub-optimal and would increase the miss-rate when evaluation is performed at high overlap thresholds; 3) The average precision measured over a range of overlap thresholds would drop when a high NMS threshold is utilized.

Considering the above situations, it seems that utilizing a decaying score of other detection boxes which has an overlap with M seems to be a promising approach for improving the detection performance. It's also clear that scores for detection boxes which have a higher overlap with M should be decayed more, as they have a higher likelihood of being false positives. Considering different decay functions with different suppressing abilities, in this paper, sigmoid derivative function will be taken into consideration.

$$s_i = s_i\left(\frac{1}{\left(1+\mathrm{e}^{(-iou(M,b_i))}\right)}\right)\left(1-\frac{1}{\left(1+e^{(-iou(M,b_i))}\right)}\right), \forall b_i \notin D \tag{1}$$

The sigmoid derivative function has a similar but fixed shape distribution, where the D represents the set of the maximum scores of the predicted bounding boxes.

3. Experiments and Results

3.1 Dataset description and experimental setup

In this paper, a public multi-class geospatial object detection dataset—NWPU VHR-10 dataset is utilized to test the performance of the proposed algorithm. This dataset contains a total of 800 optical remote sensing images, where 715 color images were acquired from Google Earth with 0.5 to 2m spatial resolution, and 85 pan-sharpened color infrared images were acquired from Vaihingen data with 0.08 spatial resolution. NWPU VHR-10 contains two datasets: a positive image set including 650 images with each image containing at least one target to be detected and a negative image set including 150 images without any targets of the given object classes. For the positive image set, 757 airplanes, 302 ships, 655 storage tanks, 390 baseball diamonds, 524 tennis courts, 159 basketball courts, 163 ground tracking fields, 224 harbors, 124 bridges, and 477 vehicles were manually annotated with bounding boxes used for ground truth. The positive image set was divided into 20% for training, 20% for validation, and 60% for testing.

The parameters of the proposed algorithm are set as below. The initial learning rate of the first-stage RPN is set as 0.001 with a "step" strategy of gamma as 0.1 and the step size as 60000; the initial learning rate of the first-stage Fast R-CNN is set as 0.001 with a "step" strategy of gamma as 0.1 and step size as 30000; the initial learning rate of the second-stage RPN is set as 0.001 with a "step" strategy of gamma as 0.1 and step size as 60000; the initial learning rate of the second-stage Fast R-CNN is set as 0.001 with a "step" strategy of gamma as 0.1 and step size as 30000. The momentum and weight decay are set

as 0.9 and 0.0005, respectively. The total iteration number of the two stage RPN are as 40000, and the total iteration number of the two stage Fast R-CNN are set as 40000.

3.2 Experimental results

To quantitatively evaluate the detection performance of the proposed algorithm, the precision-recall-curve and average precision (AP) are adopted.

	BoW	**SSCBoW**	**FDDL**	**COPD**	**Transferred CNN**	**Pre-trained Faster R-CNN (NMS)**	**Proposed algorithm (sigmoid derivative decay)**
Airplane	0.025	0.506	0.292	0.623	0.661	0.803	**0.884**
Ship	0.585	0.508	0.376	0.689	0.569	0.681	**0.723**
Storage tank	0.632	0.334	0.770	0.637	**0.843**	0.359	0.415
Baseball diamond	0.090	0.435	0.258	0.833	0.816	**0.906**	0.904
Tennis court	0.047	0.003	0.028	0.321	0.350	0.715	**0.719**
Basketball court	0.032	0.150	0.036	0.363	0.459	**0.677**	0.625
Ground track field	0.078	0.101	0.201	0.853	0.800	**0.892**	0.878
Harbor	0.530	0.583	0.254	0.553	0.620	0.769	**0.789**
Bridge	0.122	0.125	0.215	0.148	0.423	0.572	**0.611**
Vehicle	0.091	0.336	0.045	0.440	0.429	**0.646**	0.609
Mean AP	0.2457	0.308	0.245	0.546	0.597	0.702	**0.716**

From table 1, it can be seen that the proposed improved bounding box post-processing algorithm with Faster R-CNN obtains better object detection results than the traditional object detection methods as well as the pre-trained Faster R-CNN framework with NMS algorithm. For the classes of airplane, ship, bridge, the AP values obtain obvious promotion.

4. Conclusions

In this paper, an improved bounding box post-processing algorithm with Faster R-CNN framework with ZF network structure has been proposed to perform the multi-class geospatial object detection for the HSR imagery. The pre-trained Faster R-CNN framework shares the robust RPN convolutional features with the detection network, which provides a unified framework for object detection. The proposed bounding box post-processing algorithm decays all the scores of the predicted bounding boxes instead of deleting the maximum score. In addition, pre-trained Faster R-CNN utilizes the advantages of the pre-training mechanism for dealing with the limited amount of labeling information of the HSR imagery. Experimental results demonstrate that effectiveness of the proposed improved bounding box post-processing algorithm with Faster RCNN over the traditional object detection methods as well as the Faster R-CNN with NMS.

References

[1] S. Aksoy, I. Z. Yalniz, and K. Tasdemir, "Automatic detection and segmentation of orchards using very high resolution imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3117–3131, Aug. 2012.

[2] F. Zhang, B. Du, L. Zhang and M. Xu, Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection, *IEEE Trans. Geosci. Remote Sens.* , DOI:10.1109/TGRS.2016.2569141.

[3] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[4] P. Zhong and R. Wang, "A multiple conditional random fields ensemble framework for urban area detection in remote sensing optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3978–3988, Dec. 2007.

[5] N. Yokoya and A. Iwasaki, "Object detection based on sparse representation and Hough voting for optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2053–2062, May 2015.

[6] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010.

[7] Y. LeCun, Y. Bengio, G.E. Hinton, " Deep Learning," *Nature*, vol. 521, pp. 436-444, 2015.

[8] G. Cheng, J. Han, "A survey on object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11-28, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Proc. Conf. Adv. Neural Inform. Process. Syst.*, pp. 1–9, 2012.

[10] R. Girshick, 2015. Fast R-CNN. *in Proc IEEE Int. Conf. Computer Vision*, pp. 1440-1448, 2015.

[11] R. Girshick, J. Donahue, T. Darrell, J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach Intell.*, vol. 38, pp. 142–158, 2016.

[12] S. Ren, K. He, R. Grishick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," arXiv: 1506.01497v3 [cs.CV], Jun. 2016.

[13] X. Han, Y. Zhong, and L. Zhang, "An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery," Remote Sens., vol. 9, no. 7, 666; doi:10.3390/rs9070666, 2017.