PAPER • **OPEN ACCESS**

Saliency-based End-to-end Target Detection Model in Optical Remote Sensing Images

To cite this article: Fengan Zhao *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **490** 042011

View the article online for updates and enhancements.

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 171.35.171.32 on 19/09/2019 at 23:50

IOP Conf. Series: Materials Science and Engineering **490** (2019) 042011 doi:10.1088/1757-899X/490/4/042011

Saliency-based End-to-end Target Detection Model in Optical Remote Sensing Images

Fengan Zhao*, Xiaodong Mu, Peng Zhao and Zhou Yang

Department of Computer Science and Engineering, Xi'an Research Institute of High-Tech, Xi'an, China

*Corresponding author e-mail: zhao_flying123@163.com

Abstract. It is challenging work that detecting target such as aircraft in remote-sensing images because of the complicated background. Most existing methods are based on deep-learning network considering its high learning power of the features. However, the region proposal network is often based on deep network, which cost much time and computation on the proposal of the irrelevant regions which are useless to the target detection. Based on the above considerations, a novel end-to-end aircraft detection model based on saliency map is proposed in this paper. The saliency-based region proposal network can produce the target-like regions and filter out the most irrelevant background regions. Meanwhile, it cost less computing time compare to the network based on deep-learning network. Then, a novel target detection network is designed to extract the feature of target-like regions, and classify these features by the iterations of a coupled networks, the result of the binary classification is conducted by the classification layer, at same time the accurate bounding boxes are conducted by the regression layer. The performance of our method is evaluated by detecting aircraft targets in high resolution remote-sensing images. Superior experimental result proves the effectiveness and efficiency of proposed model.

1. Introduction

With the rapid development of sensor technology and remote-sensing technology, a huge number of the high-resolution remote-sensing (HRRS) images with high quality can be acquired. These HRRS images provides the researchers a chance of analyzing the semantic meaning contain in the image. Aircraft recognition is a one of the tasks of the automatic semantic interpretation, and has applied in military and many other fields. It is challenging owing to the complicated backgrounds around the aircrafts.

Up to now, most aircraft detecting algorithms can be seemed as a binary classification problem. Considering the classification principle, Aircraft detection methods can be divided into three categories. The first one is based on low-level statistical features [1,2], such as shape feature, rotation and scale invariant features, and so on. These features are manually designed, and can grasp the statistical characteristics of the aircraft. In [1] Hu moment invariant features are extracted from HRRS binary images, and can distinguish several kinds of aircraft.

Template matching methods belongs to the second aircraft detecting category. It is the early approach for target detection owing to its simplicity. Due to the types, shape and size of each aircraft is limited, researchers adopt template to represent the aircrafts which share the similarity measurement.


Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

In [3], experiments are conducted with the test dataset of 1925 aircraft images, which are from 11 types of aircrafts. These aircraft templates are elaborately designed.

The third category of aircraft detection is based on deep network. Recently, with the rapid growth of deep learning, most target detection frameworks adopt deep neural network, i.e., the convolutional neural network (CNN), because of the tremendous feature representative power. These CNN-based target detection methods can be concluded into two classes: using or not using the region proposal network. The former CNN-based detection methods adopt the region proposal network (RPN), like RCNN [4] and Faster RCNN [5]. This kind of method firstly selects the region of interests (RoIs), which are extracted from the HRRS images by selective search [6], edge box [7], or RPN [5]. Then, CNN network is adopted to extract the feature for every RoI. Finally, the bounding box which denotes the location of the targets is acquired by softmax classifier and regression layer. The latter CNN-based detection methods do not using the RPN for detection, such as YOLO [8] and SSD [9]. The YOLO or SSD model directly conduct convolution operation on the HRRS image, and do not need to use the region proposal, so it is the processing time is shorter than Faster-RCNN, but the accuracy is not satisfied. Concisely, the performance of the former detection methods with RPN is better than that of the latter without RPN. However the operation speed of the former is slower than the latter.

In this work, a saliency-based end-to-end target detection model is proposed to tackle the deficiencies of the above methods. Considering that saliency method can provide fewer region proposals, which can be seemed as a previous coarse selection of target region, we design a new aircraft detection framework. Firstly, data augmentation is adopted for enhancing the robustness. Then a saliency-based region proposal network is used to extract the target-like region, and filter out the irrelevant background regions. After that, a novel target detection framework is designed to extract the multi-level feature of target-like regions produced by the saliency-based region proposal network. The result of the binary classification and the accurate bounding boxes are conducted by the classification layer and the regression layer of the target detection network. The performance of our method is evaluated by detecting aircraft targets in optical HRRS images. The results of elaborately designed experiments show the superior detection accuracy and efficiency.

There are three main contribution of our proposed model over the other detecting methods. Firstly, a novel saliency-based region proposal network is proposed to extract the target-like region, and filter out the irrelevant background regions. Secondly, an effective end to end target detection framework is proposed to detect aircraft automatically. The framework contains a residual network which can extract powerful feature representation by feature fusion, and an alternating detection network with hard negative mining. Thirdly, a new hard negative mining (HNM) method with an iterative coupled network is proposed which enhancing the classification power and optimizes the region locations.

2. Method Description

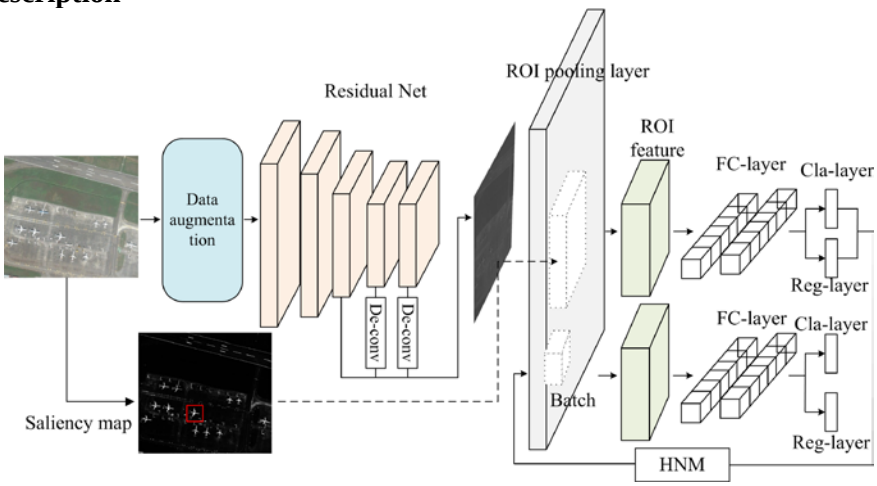

Figure 1. Proposed aircraft detection framework.

The framework of our aircraft detection algorithm is illustrated in Figure 1. The algorithm comprises three parts: data augmentation, saliency-based region proposal network and target detection network. Data augmentation is adopted to increase the data and make the algorithm more robust. The saliency-based region proposal network produces target-like proposals based on saliency detection method. The target detection network using coupled CNNs for large-scale HRRS image target detection. It takes an image with or without target as input and outputs the category (aircraft or background) of the target-like region and the possibility of each target-like region.

2.1. Data augmentation

The available HRRS images with manual annotation of object for training deep-learning models are insufficient. This is because that the HRRS image annotation conducted by human is expensive and tedious. In this work, data augment of the training dataset is employed, and it can significantly improve the robustness of the method. We randomly scale and rotating the HRRS images in hue-saturation-value (HSV) to augment the data set. Due to the different perspectives of imaging conditions, affine transformations are performed to data augmentation. Our comprehensive experiments (see Section 3.1) have shown that the randomly scaling, rotation and affine transformation can boost the performance remarkably.

2.2. Saliency-based region proposal network

It is well known that saliency model has been applied in many visual tasks due to its powerful ability of quickly accessing to information. However, for the aircraft contained in the HRRS images, the object is very small and the structure is similar to the background. The common saliency-based method fails to generate saliency image with high-quality because of the small size and the complicated background of the aircraft. Besides, the extracted region of interest (ROIs) are mostly negative samples. This can make the detection efficiency drop much, so the detection accuracy is not satisfied. So we need to add some conditions to reduce the extraction the of some irrelevant background regions. Inspired by [10], we generate the saliency map image by imposing a limit to the pixel number of the target-like regions. We found that in the HRRS image which contains the aircraft target connect less to the image boundaries than the background object does. So in this work, we can obtain salient objects by limit the pixel number of the target-like object. Specifically, at first, the ratio of the pixel number of the part at image border to the pixel number of the whole field of the object is obtained. Then pixels are replaced with super pixels [11], and the similarities of the nearby super pixels are constructed according to the calculation of the shortest path.

2.3. Target detection network

After generating the RoIs with the saliency-based region proposal network, these RoIs are put into the target detection network for detection. We first use a feature fusion method to extract the features of RoIs with a deep learning model, ResNet-50. Then after the pooling stage by the RoI pooling layer, we obtain a feature vector according to the target-like region. Lastly, the target-like region is classified into the category of target or non-target with classification layers. The bounding box of the target is regressed with regression layers. These two steps are all conducted by a coupled network.

2.3.1. Residual net

The architecture of our proposed method is based on ResNet-50 [12] because it is a deep and powerful framework, which is widely used in visual task. This network is pre-trained on the well-known ImageNet, and contains strong feature representation power. As the middle several convolutional layers of the ResNet contains structure information and the latter several convolutional layers contains more semantic information, in this work, we fuse the structure and semantic information by map combination. Specifically, we seem the middle layer as reference scale, and the last two layers with deconvolution are combined with the middle layer, to constitute a concatenated feature map.

2.3.2. Alternating detection network with hard negative mining

After the concatenated feature maps which contains different level of information are extracted, a simple yet effective alternating detection network with hard example mining is proposed. We adopted two same detecting networks: network A and network B, which have the same initial parameters. When the concatenated feature map is input to the ROI pooling layers, respective candidate regions according to the saliency map are generated. They are first input to the network A, which involves only forward operations, for the calculation of loss value. For each candidate target-like region, we obtain the respective loss value. Then, hard example mining [13] is used in model training to improve the discrimination power, and it can also improve the training effectiveness of the networks. In this work, we use hard example mining to sort the loss values of all the candidate regions.

The HNM method is simple but effective way to adding the background samples for enhancing the robustness of our framework. As the complexity of the background in the HRRS images, to cover all the background categories as much as possible with HNM is significant for the detection result. Here an iterative process is adopted for the model training. Those archives the lowest scorers are selected and output into B for training. The difference between A and B in training is that, regions which are input into network A are all the candidate regions, while the regions input to the network B are the sorted regions, which scores the worst in network A according to the loss values. The hard example which is acquired in the last iteration is added to the negative data set. The training process eventually converged when accuracy stops increasing and become stable.

3. Experiment and Results

3.1. Data and Experimental setup

To evaluate our method, 2500 HRRS images which contain the aircrafts from Google Earth are sampled. To address the difficulties in data collection of aircraft, positive aircraft samples are augmented by scale and rotation transforms. This data augmentation can also increase the diversity of the HRRS images and make the framework learn rotational features sufficiently. Comparison of the error rate with and without the data augmentation of our framework is shown in Table 1. From the result we can see that the data augment has a great influence on performance enhancement.

Table 1. Comparison of the error rate with and without the data augmentation.

Data process	**Using data augmentation**	**Not using data augmentation**
Error rate	0.039782764	0.146778323

Our HRRS image resolution is from 400 × 400 to 1000 × 1000 pixels. The ground truth of aircraft images was manually annotated. The momentum and weight decay are set to 0.9 and 0.0001 respectively in training. As for the learning rate, we adopt gradual learning rate according to the iteration number, which means learning rate is set to 0.001 in the initial 40k iterations, and 0.0001 in the following 30k iterations. Simulation is done with Lenovo K500 Gragh Workstation with Intel E5 CPU, 16GB RAM, and a GeForce K2200 GPU with 4GB RAM.

3.2. Evaluation Remarks of Simulations

Among the reference index of target detection, the precision ratio and recall ratio are the two most important criterions in the assessment of target detection. They can be denoted as:

$$Precision = \frac{N_{DS}}{N_{DS} + N_{DF}} \tag{1}$$

$$Recall = \frac{N_{DS}}{N_{TS}} \tag{2}$$

where total real aircrafts number and the real detected aircrafts number are denoted as N_{TS} and N_{DS}. The number of falsely detected aircrafts are represented as N_{DF}. The precision rate is set to maximize

the number of detected target without missing some of the targets. The recall rate is the ratio of detected targets and the whole real targets.

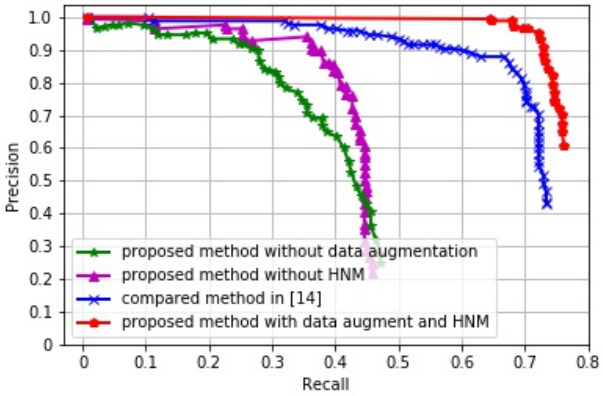

Figure 2. Precision-recall curves of different methods.

In order to evaluate the influence of the using of data augmentation and HNM on the proposed model, we make comparison among our proposed method with data augmentation and HNM, our method without data augmentation and our method without HNM. In addition, the method in [14] is also used as comparison to evaluate the proposed model. The comparison results are shown in Figure 2. From the comparison curve we notice that after augmenting of training data and hard negative mining, the performance is significantly enhanced. Furthermore, effect of hard negative mining is greater than that of data augmentation.

Figure 3. The detection result of two HRRS images from Google Earth. (a),(c) ground truth; (b),(d) detection result obtained by our proposed method. Green, red and blue bounding box means ground truth, true positive and false positive.

Figure 3 shows the test results obtained by our proposed method. Ground truth of the aircrafts is shown with green bounding boxes, true positive and false positive is in red and blue, respectively. From the results we can see that our method can accurately locate the aircrafts in the HRRS images with complicated backgrounds.

The average detecting time for each HRRS image is 2518.32 ms. Due to the simple and effective course detection procedure of saliency-based detection, most irrelevant part in the test image has been filtered out, this guarantees balance of the detection speed and good detection performance.

4. Conclusion

A novel end-to-end method to detect aircraft in optical HRRS images is presented. Our method provides a new way to detect aircraft. In the region proposal stage, we adopt saliency-based model to extract the target-like regions, and filter out the irrelevant regions which are useless to the target detection. In the target detection stage, the target-like region is represented with a feature vector, which is constructed by ResNet and a RoI pooling layer. Then target-like region is identified to target or non-target with the classification layers of a coupled network, and the bounding box of the target is

constructed with the regression layers. Experiment results show the excellent detection performance whether the target background is complicated or not.

References

[1] S.A. Dudani, K.J. Breeding and R.B. McGhee, Aircraft identification by moment invariants, IEEE Trans. Comput. 100 (1977) 39-46.
[2] X. Bai, H. Zhang and J. Zhou, VHR object detection based on structural feature extraction and query expansion, IEEE Trans. Geosci. Remote Sens. 52 (2014) 6508-6520.
[3] A. Zhao, K. Fu, S. Wang, J. Zuo, Y. Zhang, Y. Hu and H. Wang, Aircraft recognition based on landmark detection in remote sensing images, IEEE Geosci. Remote Sens. Lett. 14 (2017) 1413-1417.
[4] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition 2014 pp. 580-587.
[5] S. Ren, K.He, R. Girshick and J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 6 (2017) 1137-1149.
[6] J.R. Uijlings, K.E. van de Sande, T. Gevers and A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2013) 154-171.
[7] C.L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in Proc. of the 13th European Conference on Computer Vision (ECCV) 2014 pp. 391-405.
[8] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 2016 pp. 779-788.
[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu and A.C. Berg, "SSD: Single shot multibox detector," in European conference on computer vision (ECCV) 2016 pp. 21-37.
[10] G. Hu, Z. Yang, J. Han, L. Huang, J. Gong and N. Xiong, Aircraft detection in remote sensing images based on saliency and convolution neural network, EURASIP Journal on Wireless Communications and Networking, 1 (2018) 26.
[11] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 2274-2282.
[12] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 2016 pp. 770-778.
[13] J.F. Henriques, J. Carreira, R. Caseiro and J. Batista, "Beyond hard negative mining: Efficient detector learning via block-circulant decomposition," in proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013 pp. 2760-2767.
[14] W. Zhang, X. Sun, K. Fu, C. Wang and H. Wang, Object detection in high-resolution remote sensing images using rotation invariant parts based model, IEEE Geosci. Remote Sens. Lett. 11 (2014) 74-78.