Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 650–656

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Orientation Robust Object Detection in Aerial Images Based on R-NMS

Qing Qing Liu[a], Jian Bin Li[a,*] *

[a]*School of Information Science and Engineering, Central South University, Changsha 410012, China*

Abstract

Object detection in aerial images is a challenging task which plays an important role in many fields, such as intelligent traffic management, fishery management and so on. Different from object detection in natural images, the orientation of objects in aerial images is arbitrary. The axis-aligned bounding box detection, which is always used in traditional object detection methods, will cover a lot of redundant information and deteriorate the detection results when it is used to locate the object in aerial images. Therefore, traditional object detection methods are no longer applicable for aerial images. In order to promote the object detection performance in aerial images, we propose a novel orientation robust object detection model based on rotated non-maximum suppression (R-NMS). In addition, we adjust the anchor setting according to the diversity shapes of the aerial objects to enhance the performance of the model. Our model is tested on the public DOTA dataset, and the mAP is 16.31% higher than the baseline, indicating that our method is very effective and competitive in the object detection of aerial image.

Keywords: Object detection; aerial images; orientation robust; rotated non-maximum suppression (R-NMS)

1. Introduction

In recent years, inspired by deep learning, significant progress has been made in object detection. By learning the in-depth presentation of the region of interest (RoI), the deep learning-based detectors make it possible to precisely

* *Corresponding author: Tel. +(86)15084707224
Email: 625071211@qq.com ;

10.1016/j.procs.2019.06.102

locate and classify objects on the image. There are many works, such as R-CNN1, Fast R-CNN2, Faster R-CNN3, YOLO4, YOLO90005 and so on, have achieved excellent performance on object detection task in natural scenes. However, these methods are horizontal region-based detection methods and are not suitable for aerial image detection tasks because of the arbitrary rotation characteristics of aerial objects. Using axis-aligned bounding boxes to locate tilted objects will cover many redundant areas (i.g., background or adjacent objects), especially in dense object scenes as shown in Fig.1(a). In this case, axis-aligned bounding boxes are not conducive to the processing of non-maximum suppression (NMS) and are prone to missing detection. Therefore, although many previous methods achieve the state-of-the-art in natural scene, they are not quite suitable for object detection in aerial images.

In order to detect arbitrary-oriented objects in the aerial image, scholars have also proposed many methods. Chen, Gong, et al. propose a method which contains two stages6. First, a detection network is utilized to initially detect buildings in images. Then an orientation classifier network is applied to learn the rotation angle of the buildings. Zhang, Sun, et al. authors firstly segment the image into several areas and extract some RoIs (region of interest)7. Then, they detect the bridge in the RoIs through feature extraction and a neural network called Hopfield. Though these methods are successful in some applications, this kind of two-step approach is relatively more cumbersome and time-consuming than the end-to-end approaches. Some researches put forward end-to-end models8,9,10,11, but these models only recognize a single object in specific problems with simple datasets. And the method proposed in [8] and [9] still uses horizontal boxes.

We present a novel end-to-end model for detecting multiple categories of arbitrary-oriented objects in aerial images. The model adopts a variety of anchor aspect ratios to generate proposals that can better fit multiple categories of objects. In addition, our model eventually generates the rotated bounding box, which is shown in Fig.1 (b). This kind of minimum circumscribed rectangle reduces the interference of redundant noise during detection. Finally, post-processing with R-NMS to obtain the final detect results.

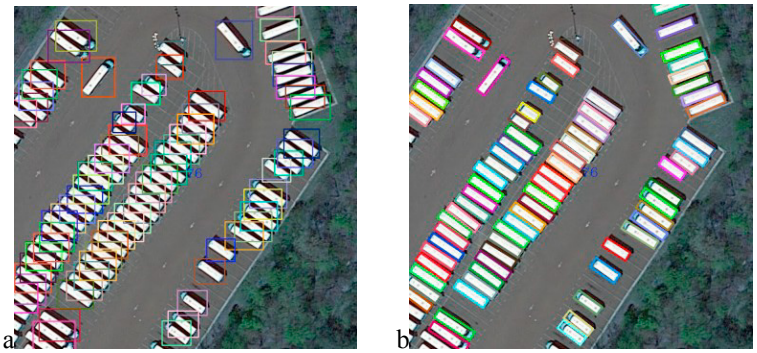

Fig. 1. Two styles of bounding boxes. (a) shows the axis-aligned bounding boxes, which covered redundant regions are relatively large. (b) shows the rotated bounding boxes, which fit object snugly.

2. Proposed Approach

2.1. Overview

Our model is based on the Faster R-CNN pipeline, its architecture as shown in Fig.2. It can be known from [2] that the Faster R-CNN is composed of two parts: regional proposed network (RPN) and Fast R-CNN. We optimized these two parts to adapt to the object detection for aerial images. The workflow of the model is as follows: First, we input the image and obtain the feature map through ResNet10112. Second, the feature map is input to the RPN to get the horizontal proposals. A variety of aspect ratios are set for the anchors in RPN to better adapt to different kind of objects in aerial images. After RPN, the model learns to classify objects and refine inclined boxes through RoI

pooling and full connected layers (FCs). Finally, in the post-processing stage, we use R-NMS to refine the object detection results. In the following sections, we will introduce the various parts of the model in detail.

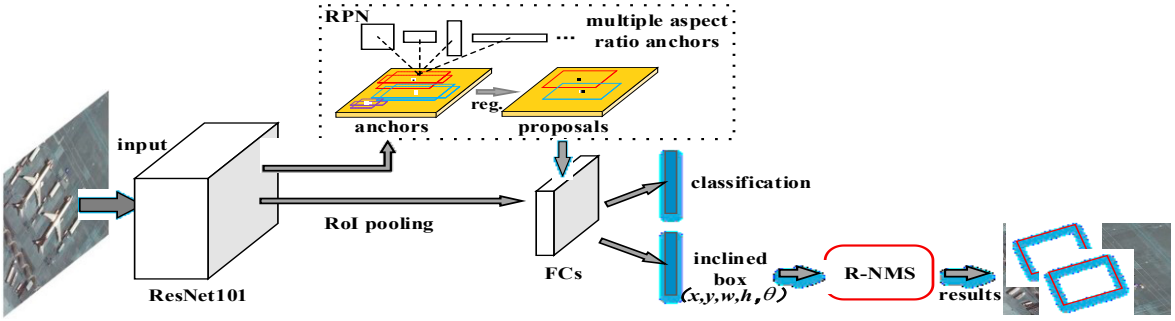

Fig. 2. Model architecture

2.2. Multiple anchor aspect ratios

The Faster RCNN sets the anchor ratios to [1, 1/2, 2]. Such setting can well cover almost everything in the nature scene. However, unlike in natural scene, aerial images 'look' at objects from high altitude and many objects often have large aspect ratios, such as the harbor, bridge and so on. According to the statistics characteristic of the objects in the DOTA13, the aspect ratio of objects in the aerial images mainly distribute from 1 to 5. So, we adjusted the aspect ratio setting to [1, 1/2, 2, 1/3, 3, 1/4, 4, 1/5, 5]. Such setting can better cover the different classes of object and getting as much information of the object as possible for further operation.

2.3. Rotated bounding box

We use five parameters (x, y, w, h, θ) to represent the rotated bounding box, which represent the coordinates of the centre point, width, height, and rotation angle of the bounding box, respectively. As shown in the Fig.3, the rotation angle θ is the angle that the horizontal axis (x-axis) rotates counterclockwise when it encounters the first edge of the box. The range of angles defined as [-90,0).

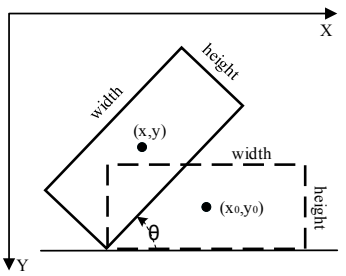

Fig. 3. The representation of rotated bounding box

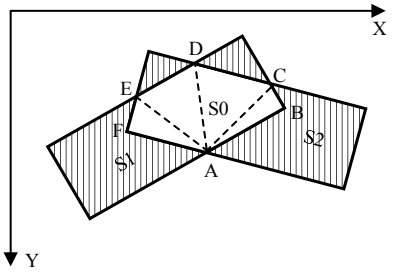

Fig. 4. Intersection area between two rotated rectangles

2.4. Rotated non-maximum suppression(R-NMS)

The core idea of NMS (Non-Maximum Suppression) is to calculate the Intersection-over-Union (IoU) between two rectangles. Because the IoU computation on axis-aligned boxes may lead to an inaccurate IoU of skew interactive bounding boxes. We design the R-NMS, which calculates skew IoU by using a triangulation method. Fig.

4 shows the geometric principles. Firstly, we should divide the intersection area into triangles and summarize all the area of the triangles (S0). Then the IoU can be obtained by using the area of two intersecting rectangles S1 and S2. The calculation formula for IoU is as follow:

$$IoU = S0/(S1+S2-S0) \tag{1}$$

2.5. Loss function

Our loss function for an image is defined as follow 2:

$$L(p,l,u,u^*) = L_{cls}(p,l) + \lambda[l \geq 1]L_{loc}loc(u,u^*) \tag{2}$$

$p = (p_0, p_1, ..., p_k)$ represents a discrete probability distribution of the k+1 categories calculated by the SoftMax function. $l = 1,2,...,k$ denotes the label of each class, and the background is labeled as 0. $L_{cls}(p,l) = -\log(pl)$ is the log loss for class l. $L_{loc}(p,l)$ is a smooth L1 loss function2 for position regression. The Iverson brackets $[l \geq 1]$ indicates that l is 1 when l>1, otherwise 0. So, there is no position regression for background. The trade-off between two terms are controlled by λ. In our model, λ is set to 1. $u = (u_x, u_y, u_w, u_h, u_\theta)$ represents the predicted tuple of the rotated bounding box, and $u = (u_x^*, u_y^*, u_w^*, u_h^*, u_\theta^*)$ denotes the ground truth tuple of the rotated bounding box. In order to facilitate the regression of the bounding box, these parameters are scale-invariantly parameterized as follow3:

$$u_x = \frac{x-x_a}{w_a}, u_y = \frac{y-y_a}{h_a}, u_h = \log\frac{h}{h_a}, u_w = \log\frac{w}{w_a}, u_\theta = \theta - \theta_a \tag{3}$$

$$u_x^* = \frac{x^*-x_a}{w_a}, u_y^* = \frac{y^*-y_a}{h_a}, u_h^* = \log\frac{h^*}{h_a}, u_w^* = \log\frac{w^*}{w_a}, u_\theta^* = \theta^* - \theta_a \tag{4}$$

where x, y, w, h represent the central coordinates, length, and width of a box, respectively. x, x_a, x^* are for the predicted box, anchor box, and ground truth box, respectively. The same applies to y, w, h, and θ. And the θ belongs to the rotated bounding box only.

3. Experiments and analysis

3.1. Dataset and training

We comprehensive evaluate our method on DOTA[13], which contains 2806 aerial images collected from different sensors and platforms. DOTA totally annotated 188,282 instances and covered 15 common categories (including the plane, ship, storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor, bridge, large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), soccer ball field (SBF), and swimming pool (SP). Each instance is labeled with an arbitrary quadrilateral.

We build our model with tensorflow and use ResNet101 to initialize the network. All models are trained and tested on an Nvidia Titan Xp GPU with 12GB memory. Before the training, we randomly flip images and subtract the mean value [103.939, 116.779, 123.68].

3.2. Experiments

Table1 summarizes the effectiveness of different settings of our model. FR-H[13] is the officially provided baseline, which means Faster RCNN trained on horizontal bounding boxes and evaluated with the rotated ground truth.

Table 1. The results of ablation experiments performed on DOTA.

Method	R-NMS	Anchor aspect ratio	mAP(%)
FR-H（baseline）	N	[1,1/2,2]	39.95
Method1	Y	[1,1/2,2]	54
Method2	Y	[1,1/2,2,1/3,3,1/4,4,1/5,5]	56.26

Method1 is used to test the effectiveness of R-NMS in object detection. It can be found that using rotated bounding boxes to train the model and post-process the prediction result with R-NMS can achieve 54% mAP and dramatically higher by 14.05% than FR-H. This proves that R-NMS is a useful method in the aerial images object detection. As shown in Fig.5, the lifting of recall is especially noticeable for densely arranged inclined objects.

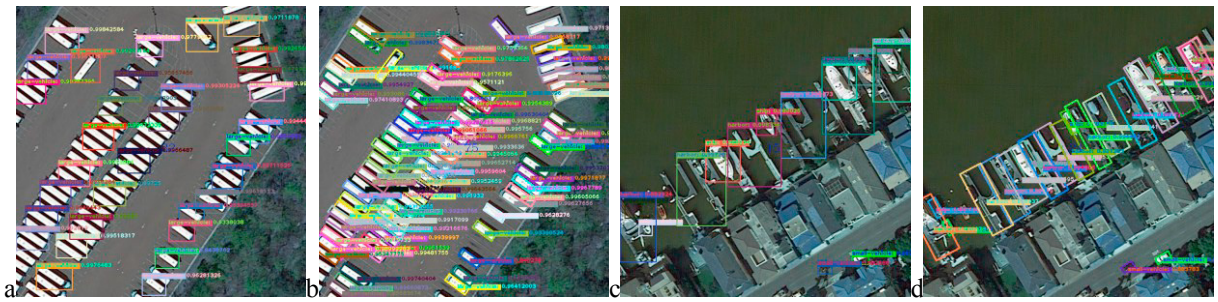

Fig 5. Detect results by using different NMS. (a) (c)Detect results obtained after the normal NMS of the horizontal bounding box. (b) (d)Detect results obtained after the R-NMS of the rotated bounding box. The numbers in the centre of the pictures indicate the number of objects detected.

In order to make the proposals to cover different categories of objects better and facilitate the further detection procedures, we set the aspect ratio to [1,1/2,2,1/3,3,1/4,4,1/5,5]. Fig.6. shows the effect of different anchor aspect ratios settings. The color boxes in the Fig.6. indicates proposals. Obviously, the proposals in (a) can cover large aspect ratios objects well, while (b) cannot. Method2 in Table 1 examines the effect of multiple anchors aspect ratios on detection performance. Its mAP is 56.26%, which is 2.26% higher than Method1. It demonstrates that utilizing multiple anchor aspect ratios can increase the flexibility of the model and improve the detection performance.

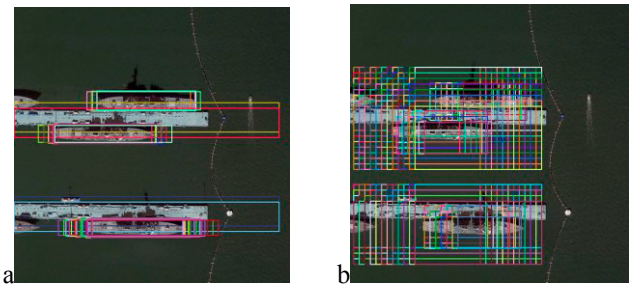

Fig 6. The effect of using different anchor aspect ratios. The colour boxes in figures indicates proposals. (a) shows the proposals obtained by setting the anchor aspect ratios to [1,1/2,2,1/3,3,1/4,4, 1/5,5]. (b) shows the proposals obtained by setting the anchor aspect ratios to [1,1/2,2]

Table 2. Detection results of different algorithm on DOTA.

	SSD[13]	YOLOv2[13]	R-FCN[13]	FR-H[13]	FR-O[13]	Our
plane	39.83	39.57	37.8	47.16	79.09	77.09
BD	9.09	20.29	38.21	61	69.12	66.14
bridge	0.64	36.58	3.64	9.80	17.17	32.69
GTF	13.18	23.42	37.26	51.74	63.49	60.23
SV	0.26	8.85	6.74	14.87	34.20	58.23
LV	0.39	2.09	2.60	12.80	37.16	46.19
ship	1.11	4.82	5.59	6.88	36.20	53.97
TC	16.24	44.34	22.85	56.26	89.19	76.86
BC	27.57	38.25	46.93	59.97	69.6	61.52
ST	9.23	34.65	66.04	57.32	58.96	71.36
SBF	27.16	16.02	33.37	47.83	49.40	45.29
RA	9.09	37.62	47.15	48.70	52.52	48.98
harbor	3.03	47.23	10.60	8.23	46.69	56.67
SP	1.05	25.50	25.19	37.25	44.80	53.38
HC	1.01	7.45	17.96	23.05	46.30	35.12
mAP(%)	10.59	21.39	26.79	36.29	52.93	**56.26**

Table 2 shows the comparisons of our method and some classical object detection algorithms on DOTA. Among them, SSD, YOLOv2, R-FCN, and FR-H are all trained on axis-aligned bounding box. FR-O denotes Faster RCNN, which is trained on rotated bounding boxes. All methods are evaluated on the rotated ground truth. As you can see, the mAP of our method is higher than other methods, which proves that our method is effective and competitive in detecting objects in aerial image. And Fig.7. exhibits more detect results of our method on DOTA.

Fig 7. Detect results of our method on DOTA.

4. Conclusion

We present a novel end-to-end model for detecting multiple categories of arbitrary-oriented objects in aerial images. The model adopts a variety of anchor aspect ratios to generate proposals that better fit multiple categories of objects. Besides, The R-NMS post-processing method can improve the recall rate. Finally, a series of experiments based on DOTA datasets prove that our method is effective.

References

1.Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).
2.Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

3.Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).
4.Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
5.Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. arXiv preprint.
6.Chen, C., Gong, W., Hu, Y., Chen, Y., & Ding, Y. (2017). Learning Oriented Region-based Convolutional Neural Networks for Building Detection in Satellite Remote Sensing Images. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 461.
7.Zhang, J., & Sun, G. (2011, December). Recognition of bridge over water in remote sensing image using discrete hopfield neural network. In Transportation, Mechanical, and Electrical Engineering (TMEE), 2011 International Conference on (pp. 439-442). IEEE.
8.Xu, Y., Yu, G., Wu, X., Wang, Y., & Ma, Y. (2017). An enhanced Viola-Jones vehicle detection method from unmanned aerial vehicles imagery. IEEE trans Intell Transp Syst, 18(7), 1845-1856.
9.Tang, T., Zhou, S., Deng, Z., Lei, L., & Zou, H. (2017, July). Fast multidirectional vehicle detection on aerial images using region based convolutional neural networks. In Geoscience and Remote Sensing Symposium (IGARSS), 2017 IEEE International (pp. 1844-1847). IEEE.
10.Tang, T., Zhou, S., Deng, Z., Lei, L., & Zou, H. (2017). Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks. Remote Sensing, 9(11), 1170.
11.Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., & Guo, Z. (2018). Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. Remote Sensing, 10(1), 132.
12.He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
13.Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]//Proc. CVPR. 2018.