



Large-Scale Target Detection and Classification Based on Improved Candidate Regions

Runing Xi^{1,2,3}, Qianqian Han^{1,2,3(✉)}, Gaoyun Jia^{1,2,3}, and Xuefeng Kou⁴

¹ School of Computer Science and Technology, Northwest Polytechnical University, Xi'an 710072, China

² National Engineering Laboratory of Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an 710129, China

³ Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, Xi'an 710129, China

⁴ Shaanxi Baocheng Aviation instrument Co., Ltd., Baoji, China

Abstract. Target detection of aerial images has become a frontier subject of concern in the image processing field. Using existing method to detect and classify large-scale building objects in aerial images, the accuracy is still a little low. This is mainly because the current method does not make full use of the prior information of the target to be detected, so there are too much redundant information in the candidate box. In this paper, our own dataset were built and then utilize the Hough transform to filter out the images that may exist in the sequence image. For images with dense lines or circles, it is possible that there is an artificial building target which will be detected, otherwise it is excluded directly. Besides, this paper exploits significance analysis from the filtered image and then extract the area of interest where the potential target is located. The results of the above-mentioned processing lay a good foundation for the subsequent detection and classification which can help improve the accuracy.

Keywords: Target detection · Significance analysis · Hough transform

1 Introduction

In recent years, one of the important topics in the field of aerial image recognition is how to obtain useful information from images efficiently and accurately. With the rapid development of science and technology, computer vision technology and drone technology have developed rapidly. Convolutional neural network (CNN)

Supported by: [1] the National Natural Science Foundation's project "Research on Multi-source Image Cooperative Detection Method Based on Biological Vision for UAV Groups" (No. 61572405). [2] Major Science and Technology Project of Shaanxi Province "Development and application demonstration of Apple's quality and safety supervision and traceability system based on the Internet of Things".

is a model of extended neurocognitive machines. The idea based on the Region-CNN (RCNN) detection method is to select the area where the target may exist from the image in advance. Then put the candidate area into the CNN to extract the features. In the end, classify and identify the object. Faster RCNN improves the candidate region generation method that can make the candidate region completed on the GPU. The result shows it can greatly reduce the detection time without affecting the detection accuracy. Its structure removes the image segmentation method extraction based on CNN. The steps of the target possible region are realized by adding a new network-RPN (Region Proposal Network) in the CNN. The overall detection process is shown in Fig. 1.

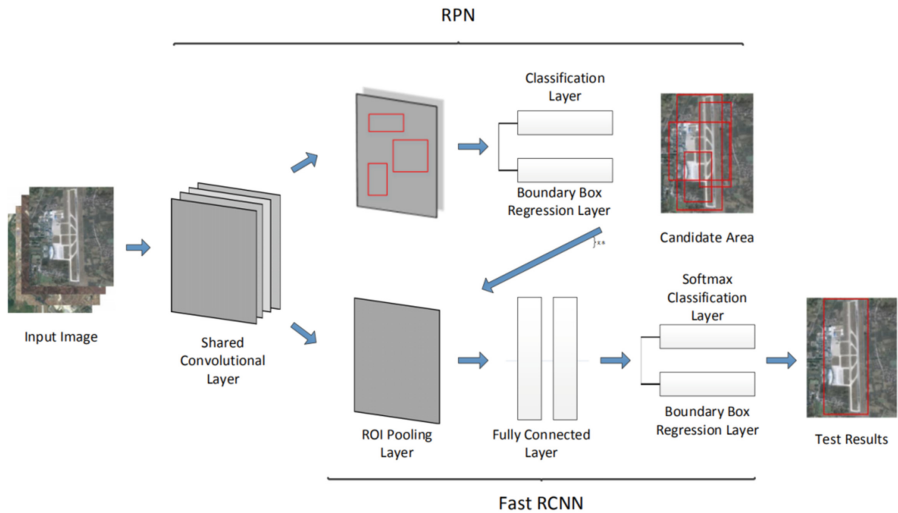


Fig. 1. Faster RCNN detection process.

Based on the Faster RCNN algorithm, we can detect, recognize and classify the target objects from the massive photos taken by UAV. But the RPN network needs to search for the candidate areas of all the pixels one by one [1-4]. This computation is relatively time-consuming. More-over, the generated suggestion box is based on the entire image and does not make full use of the apriori information of the target to be detected. Therefore, there are still too many redundant information in the candidate frame.

For the detection of large man-made structures such as airports, oil depots, bridges and so on, it is considered that these targets are of great significance in view of the great difference among the targets, the surrounding environment and the unique geometry of each target. Thus saliency detection mechanism can be introduced into the detection of these large-scale targets, which can achieve the purpose of extracting the target candidate areas. Meanwhile, it can overcome the shortcomings of the conventional methods which use the sliding window to detect

the pixels one by one, resulting in too slow speed. Otherwise, it solve the problem that inaccurate positioning of building targets in amplitude images [5–9]. After obtaining the area of interest of the potential target, the features will be extracted and classified by the Faster RCNN network, finally identify the target.

This paper use the Hough transform to detect whether there is a target as a preliminary screening, which can reduce the processing and analysis of the image without the target, and accelerate the detection speed. It is considered that there is a target to be detected when there is a dense linear and circular image, otherwise excluded directly. Then saliency analysis on these remaining images to extract the candidate areas where the potential targets are located. The flow is shown in Fig. 2.

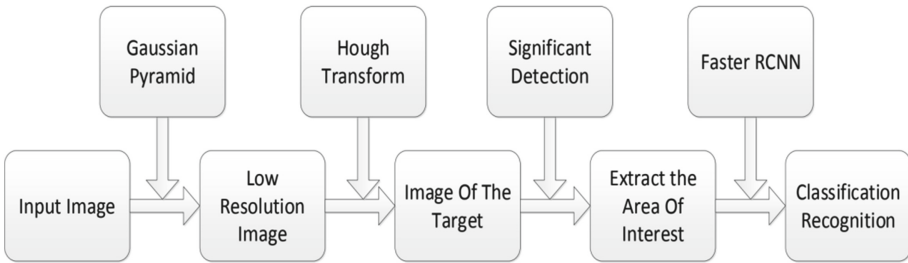


Fig. 2. Flow chart of large-scale target detection in aerial image.

2 Improved Hough Transform

The Hough transform is used to detect whether there is a target for a preliminary screening, which can reduce the processing and analysis of the image without a target, and speed up the detection speed. For the image with dense straight lines and circles, it is considered that there may be a target to be detected, otherwise it is ruled out directly.

2.1 Classical Hough Transform

Hough transform is a geometric feature extraction technology in image processing, and it is not affected by the image rotation. It transforms the image coordinate space into the parameter space, and calculates the local maximum of the cumulative result through a voting algorithm that will achieve a straight line and curve fitting, its main role is to detect from the image which has a certain characteristic of the geometric shapes, such as lines, circles and so on [10]. In Cartesian coordinates, a straight line can be expressed as an equation:

$$y = kx + b \quad (1)$$

This means that a straight line in the parameter space k - b , i.e. a point that can determine a linear cluster with parameters k , b . Based on the above theory, the binarized image can be edge-detected firstly, and each non-zero pixel point in the image is mapped to a parameter plane as a straight line. Therefore, when Hough transformation is performed on all points on the image, the detected straight line must be the point where the straight line intersects the most in the parametric plane, and all lines can be detected by traversing the entire image. However, in practical applications, since $k = \infty$ (infinite slope) cannot be represented, the Cartesian coordinate system is usually converted into a polar coordinate system. The straight line can use the parametric equation shown in Eq. 2.

$$\rho = x \cos \theta + y \sin \theta \tag{2}$$

In polar coordinate system, the method of transforming polar coordinate points into Hough space lines is shown in Fig. 3.

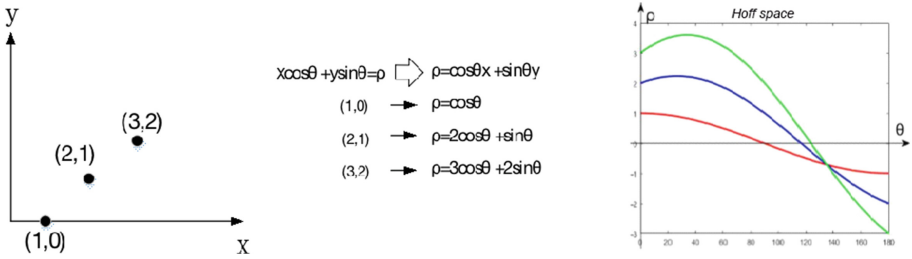


Fig. 3. Schematic diagram of converting polar coordinates to Hough space.

The Hough transform is also suitable for detecting curve whose function representation has been known. Because if the equation of the curve is known in the plane coordinate system, the corresponding parameter representation can be found and one point in the image coordinates can be mapped to the corresponding curve or surface in the parameter space. When using the Hough transform to detect a circle, the circle's space coordinate representation should be converted into a parameter space representation firstly. The general equation for the known circle is shown in Eq. 3.

$$(x - a)^2 + (y - a)^2 = r^2 \tag{3}$$

Where (a, b) is the center of the circle and r is the radius of the circle.

Convert the circle in the plane rectangular coordinate space to the parameter space, then a point (x,y) on the circle in the plane corresponds to a three dimensional cone surface with a change in height r in the parameter space, as shown in the left figure of Fig. 4. In the Cartesian coordinate system, a circle passing any point corresponds to a three-dimensional cone in the parameter space. When Cartesian coordinates are converted to the parameter space, and a

point on the same circle corresponds to a three-dimensional cone surface, it must intersect at a point (a_0, b_0, r_0) where the height is r_0 . By detecting this point, three parameters of the circle can be obtained and the corresponding circle can be found.

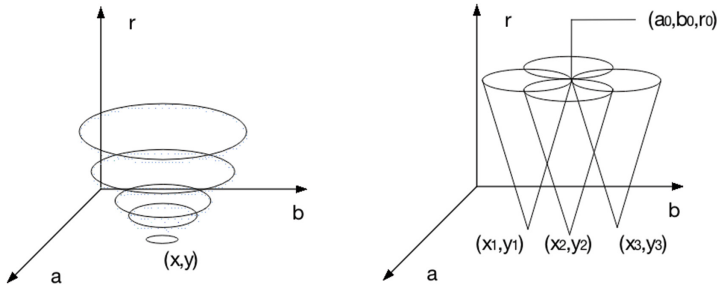


Fig. 4. The spatial representation of the parameter.

The effect of treating the oil field with the Hough transform is shown in Fig. 5. The first subfigure is the original image, and the next subfigure is the image after edge detection. The third is the image after Hough transform.

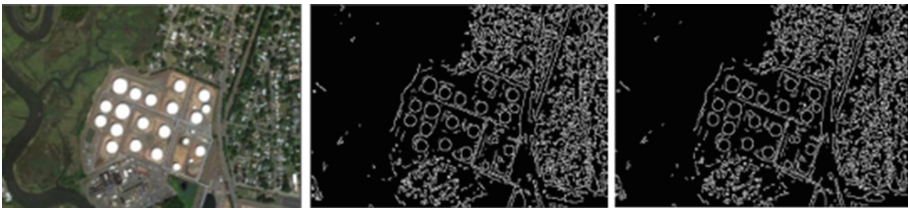


Fig. 5. Hough transform.

2.2 Improved Hough Transform

Because the parameter space of the Standard Hough transform needs three variables, obviously, the amount of calculation will be large, and the running memory and speed will be affected. The gradient direction can be used to find the center of the circle quickly and reduce unnecessary blind search in the improved method, but for most gray-scale images, the target after edge extraction still has a certain width. There are still many pixel points used to search for the center of the circle coordinates, so the problem of calculation amount has not been solved well. In this paper, we combine the Standard Hough transform and the unique properties of the circle to improve the efficiency of the operation by reducing the

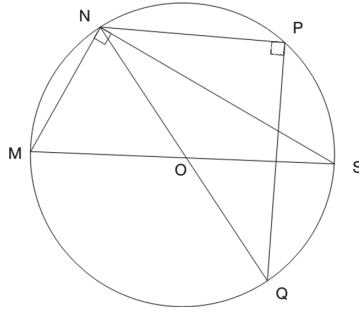


Fig. 6. Finding the center of a circle according to its characteristics.

dimension. The improved algorithm is divided into two steps: first find out the center of the circle, and then determine the radius.

As shown in Fig. 6, starting from a certain point m in the edge of a circle, according to a certain fixed direction, for example, a point N is taken out of n points clockwise, and a passing point N is taken as the vertical line of line MN . The vertical line must intersect the circle at a point S , connecting point M and point S , which is a diameter of the circle. According to the nature of the circle, the line must pass through the center of the circle. Starting from point N , a point P is taken out of n points clockwise again. Make a straight line perpendicular to the segment NP , and the straight line must intersect a point Q on the circle, connecting point N and point Q . Similarly, NQ also passes through the center of the circle. In this way, the intersection of segment MS and segment NQ must be the location of the center of the circle. According to the above principle, combined with the voting mechanism in Hough transform, the accumulator of the center point will be increased every time the center point is determined. When the search is finished, the coordinate value that reaches the preset threshold in the accumulator is the center coordinate. Then, from the non-zero pixels around the center of all the candidate circles, the radius is determined according to the degree of support for the center. Because the two-dimensional coordinate values are stored in the accumulator, compared with the Standard Hough transform, the one-dimensional is reduced, which greatly saves the storage space and calculation time. As shown in Fig. 7, this is the result pictures which might contain targets detected by improved Hough transform.

3 Significance Test

In the research of visual saliency target detection, visual attention mechanism is the focus of researchers' attention, and target saliency extraction works according to visual attention mechanism [11].

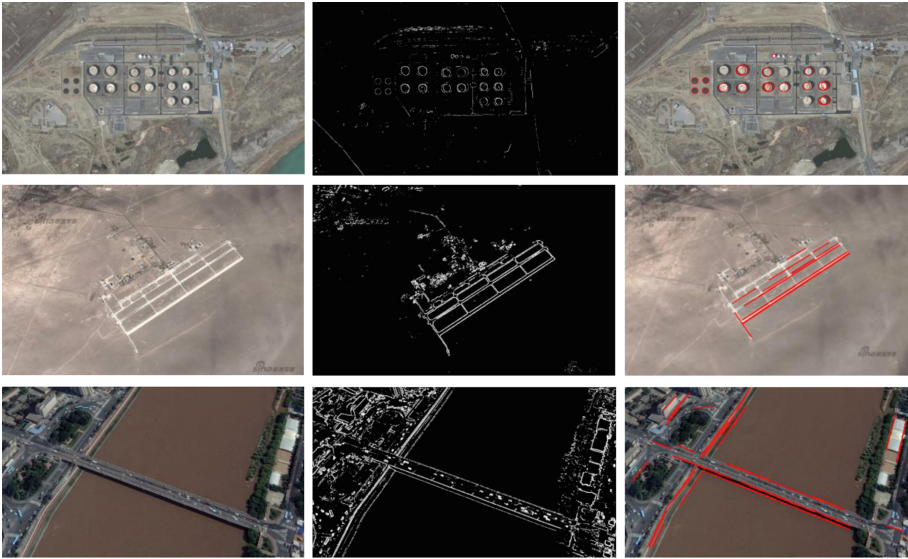


Fig. 7. Detection result by improved Hough transform.

Human visual attention mechanism can focus limited cognitive resources on the important information in the scene and suppress those relatively unimportant information. The current visual attention mechanism includes two kinds of visual models: one is the pre-attention mechanism with bottom-up control strategy; the bottom-up information processing is mainly the visual stimulation contained in the image, which has nothing to do with the high-level knowledge related to human brain tasks; the top-down information processing involves not only the prior knowledge of the target, the preparatory knowledge of the scene, but also the high-level knowledge of emotion, intention, expectation, motivation, etc.

In this paper, we use the bottom-up attention mechanism, the method of calculating visual saliency based on random sampling [12], which first samples the image into some random regions of interest, then calculates the saliency of each part on these regions, and finally fuses all regions. The flow chart is shown in Fig. 8. For an RGB image, the Gaussian filter is used to filter it first and transform it from RGB space to lab space. Compared with other color spaces, lab space is closer to human's neural visual space and is considered as the standard color space for calculating saliency image. Several windows are randomly generated in each of the three channel images because of the saliency area or objects. The volume may appear in the image in any position and size, so the randomly generated strategy helps the generated window contain all areas as much as possible, then calculate the ratio of the area of each window to the sum of gray scale, and then calculate the saliency mapping of each window and each pixel; finally, the method of Euclidean distance is used to fuse the saliency value

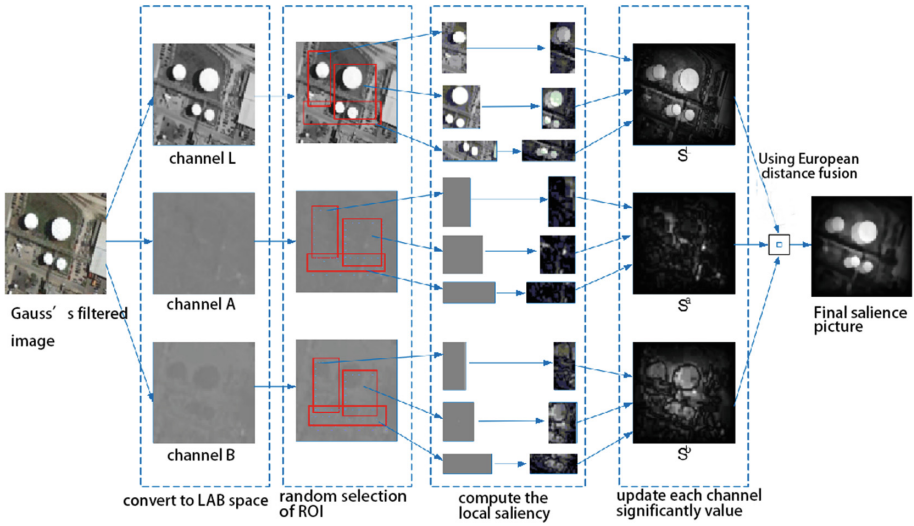


Fig. 8. Flow chart of visual significance detection based on random sampling.

of lab channel into the saliency map, and The final saliency map is generated after the value is normalized to [0,255].

The salience picture calculated by above method is shown in Fig. 9. After image segmentation, the result of candidate area generated by the method in this paper can be seen in the rightest part of Fig. 9. It is obvious that the segmentation effect is good, which can be used as the basis for subsequent classification and recognition.

4 Analysis of Experimental Results

4.1 Experimental Setting

In network structure shown in Fig. 1, this paper uses small-scale ZFNet [13] and another medium-scale VGGNet [14] as the shared convolution layer, respectively. They were run in two different experimental environments: the PC and the server built by the lab. The hardware and software environment used is shown in Table 1.

In this paper, the initial data set of airport oil depot target is constructed by using the image of Internet and the satellite image captured by Google Earth. The collected image data set includes 780 airport targets, 897 oil depot targets, 408 bridge targets, including 2085 images, covering a variety of samples in each type of typical targets. After change processing and equalization, the experimental data set contains 4680 airport targets, 5382 oil depot targets, 4496 bridge targets, including 14558 images.

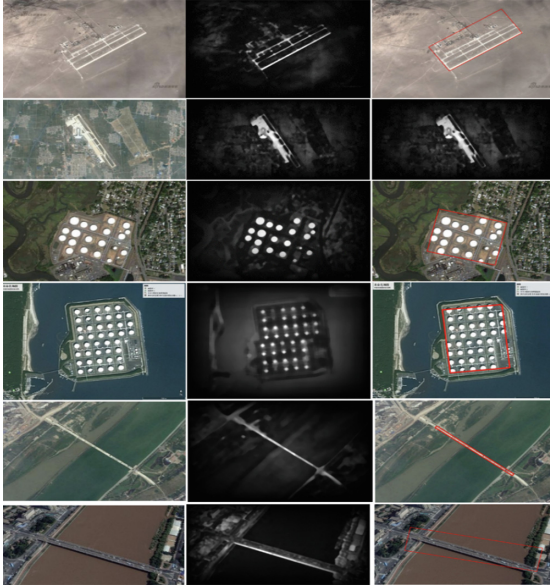


Fig. 9. Extracting candidate regions from large scale images.

Table 1. Software and hardware environment used in the experiment.

Project	ZF + Faster RCNN	VGG + Faster RCNN
Processor Inter	Inter i7-2600k 3.40 GHz	Xeon E5-2640v4 3.40 GHz
RAM	8GBz	64GB
Memory	GTX 970 (4G)	Tesla K40 (12G)
Operating System	Windows 7	centos 7
Deep learning framework	Caffe	TensorFlow
Data processing software	MATLAB R2016b	Python 3.0

4.2 Training Setting

The goal of the detection and classification contains three categories including airport, oil depot and bridge. Therefore the output of the input layer, the candidate area layer and the classifier is set to 4. The output of the bounding box regression layer is the four coordinate parameters of the bounding box, so its output is set to 16. Other parameters including the initial learning rate and the stepsize are set as shown in Table 2.

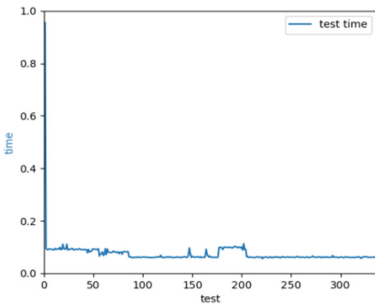
Table 2. Parameter settings of network training.

Parameter name	Value
base_lr	0.001
Gamma	0.1
Stepsize	50000
Display	20
Momentum	0.9
iter_size	2

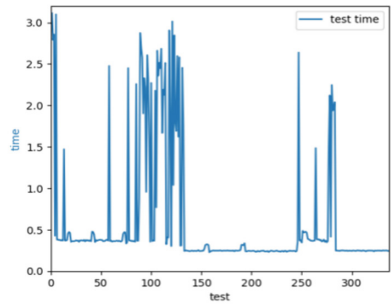
4.3 Classification Network Experiment Results

As shown in Fig. 10, the time spent in each image test in the test set is changed. For the convenience of viewing data, the running time under the two shared convolutions is displayed in the coordinate system of different scales. Under the shared convolution of ZF network, the total time consumed is 25.372 s. On average, one image is tested every 75 ms, and the processing time of each image is relatively uniform. As a whole, under the training of small-scale network ZF, classification detection can meet the real-time requirements to a certain extent. But under the shared convolution of VGG network, the total time consumed is 198.503 s, and it needs to complete an image test every 0.589 s on average. From this analysis, it can be seen that VGG network is difficult to meet the real-time requirements of classification and detection. Considering the hardware and software environments of ZF network and VGG network, it can be concluded that the time cost of ZF network is much less under the same conditions.

Consider that the SSD network [15] is also a good network for detection. Therefore, the two networks mentioned above are compared with the SSD network. For the airport, oil depot and Bridge targets in the self-built database, the accuracy of SSD network and two different shared convolution networks is



(a) ZF + Faster RCNN



(b) VGG + Faster RCNN

Fig. 10. Extracting candidate regions from large scale images.

shown in Table 3. It can be seen that compared with SSD method, the accuracy of the improved Faster RCNN algorithm is significantly improved. On the basis of VGG network with medium complexity, VGG + Faster RCNN is 10% points higher than SSD, and the effect is very obvious. And in the case of 80000 iterations, the classification detection rate of the two Faster RCNN networks is high, and the difference is only 4.56%, which can meet the requirements of classification detection.

Table 3. Accuracy of various targets in different shared convolution layers.

Scheme	Airport	Oil depot	Bridge	mAP
SSD	0.8575	0.7809	0.8335	0.8115
ZF + Fast RCNN	0.8733	0.8045	0.8537	0.8689
VGG + Fast RCNN	0.9229	0.8861	0.9106	0.9145

As shown in Figs. 11 and 12, the detection results of different classified targets in different shared convolutions are shown. It can be seen from the figure that most of the targets can be distinguished and detected by the classification algorithms under the two network structures, but for some targets whose profile information is not too obvious, ZF algorithm still has the situation of missing detection, while VGG algorithm can distinguish this well because of more network structure layers, such as ZF model in Fig. 12(d).

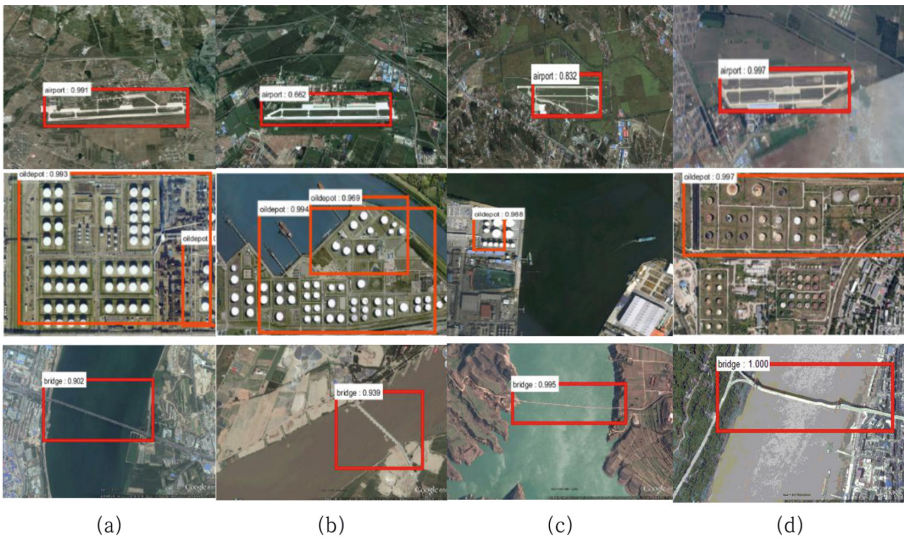


Fig. 11. Test results under ZF model.

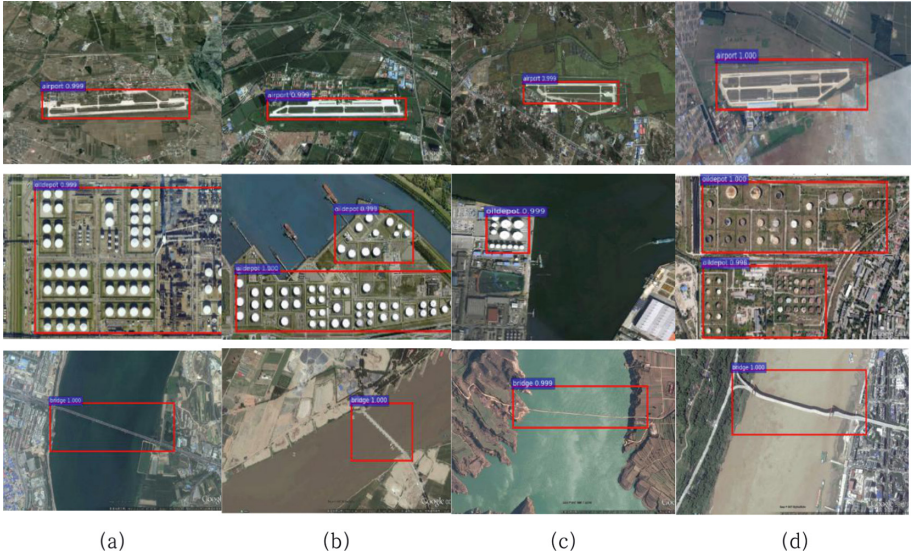


Fig. 12. Test results under VGG model.

5 Summary

In order to improve the efficiency and accuracy of detection and classification of large-scale targets such as airport et al. in aerial photos from UAV, this paper proposes a region of interest extraction strategy based on the combination of image prior structure features and visual significance detection, which can quickly find potential targets from large-scale UAV images area. Considering that the structural information of the target to be detected is the main feature for the category classification, the original image is reduced to a low-resolution image through the Gaussian pyramid firstly. Then the region of interest of the target is extracted by the combination of the prior structural feature of the image and the visual significance detection. Finally, the candidate region is generated by RPN network. The processing of Faster RCNN identifies the category of the target. The experimental results show that the ROI extraction strategy proposed can detect the potential target area in UAV aerial image quickly, eliminate a large number of redundant information in the image, and accelerate the speed of detection and classification.

References

1. Sun, W., Cheng, H., Qiu, R.: Remote sensing image target localization algorithm. *Infrared Technol.* **10**, 831–835 (2015)
2. Wang, H., Dong, Y., Yuan, B.: An effective line extraction algorithm in aerial image. *J. Wuhan Univ. (Inf. Sci. Edn.)* **37**(2), 160–164 (2012)

3. Wang, H.: Research on target recognition and tracking technology of UAV based on visual perception. Beijing Institute of Technology (2015)
4. Tong, X.: Research on moving target detection and tracking method of aerial video. Northwestern Polytechnical University (2015)
5. Sarkar, S., Duncan, K.: Relational entropy-based saliency detection in images and videos. In: IEEE International Conference on Image Processing. IEEE (2013)
6. Dapeng, L., Longsheng, W.: Saliency remote sensing image object detection model based on visual attention mechanism. *Comput. Eng. Appl.* **50**(19), 11–15 (2014)
7. Meng, L.: Saliency detection for color image based on visual attention mechanism. *Appl. Res. Comput.* **30**(10), 3159–3161 (2013)
8. Numano, S., Enami, N., Ariki, Y.: Task-driven saliency detection on music video. In: Jawahar, C.V., Shan, S. (eds.) ACCV 2014. LNCS, vol. 9009, pp. 658–671. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16631-5_48
9. Wu, F.J., Wei, C.C., Guan, S.Q.: Tool wear detection based on visual saliency mechanism. *Appl. Mech. Mater.* **602–605**, 1891–1894 (2014)
10. Leavers, V.F.: Which Hough transform? *CVGIP Image Underst.* **58**(2), 250–264 (1993)
11. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. *Vis. Res.* **40**(12), 1489–1506 (2000)
12. Lei, Y., Ji, M.: A study of the classification of imbalanced data streams based on random balance sampling. *J. Yunnan Univ. Nationalities (Nat. Sci. Edn.)* **027**(001), 63–68 (2018)
13. Trecvid, Q.A., Related, S.: Visualizing and Understanding Convolutional Networks [104]
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *Computer Science* (2014)
15. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: Single Shot MultiBox Detector. Springer, Cham (2016)