A Object detection Method for Missile-borne Images Based on Improved YOLOv3

Shaobo Wang 1,a, Cheng Zhang1,b, Di Su1,c and Tianqi Sun1,d
1 Key Laboratory of Dynamics and Control of Flight Vehicle, Ministry of Education, Beijing Institute of Technology, Beijing 100081, China
E-mail: 5shao−bo.wang@bit.edu.cn; 5,b 5zhangcheng@bit.edu.cn; 5,c 53220195008@bit.edu.cn; 5,d 5suntianqivip@163.com

Abstract. Detecting small objects in complex circumstances is an important topic in the research of today’s object detection [1], especially in military, which needs more reliable, stable and accurate detection results. In order to improve the detection of small objects, we improved the structure of the YOLOv3 network by replacing the convolution module in the original network with multi-branch scale convolution, increasing the adaptability of the network to different sizes of objects and reducing the number of network layers to balance the depth and width of the network, while also improving the feature extraction and representation capabilities. And based on the premise of a small number of data sets, we simulate some complex environments, which are composed of different weather, illumination, motion and rotational blur. We also enhance and extend the data in the network learning. Through the system simulation experiment, small objects can be recognized in such complex environments, which provides a reference for object detection of missile-borne images.

1. Introduction

With the improvement of science and technology and the modernization of war mode, the weapon system should be equipped with swift reaction speed and attack accuracy. Distant range object often appears in the form of small objects in the field of view, and the contrast between the object and the background is shallow [1], so it is more difficult to detect the object. Based on the above analysis, detecting small objects in a complex battlefield is of great military value and research significance.

In the field of deep learning object detection, how to successfully apply it to weapon system is very difficult. There are three main problems [2]: first, the amount of data is insufficient. In the military field, the most missing is the number of photos of object investigation; second, the climate in which the investigation is taken is different from that of attack, including illumination, noise and weather conditions, which fail to guarantee that the detection algorithm can adapt to the attack of the whole climate through one investigation data learning; third, it is difficult to detect small objects (The definition of small objects refers to MS COCO data set), as shown in Table 1.

Based on the above difficulties, we improve the network of YOLOv3 [3], increase the detection characteristics of small-scale objects, change the convolution module of the original network into multi-branch scale convolution, and use the data enhancement method to make the neural network detect small objects in complex environment more easily. The experimental
platform of this paper is carried out through the mountain attack scene simulated by the visual simulation software.

| Table 1. Definition of small objects. |
|--------------------------------------|
|                                       |
| Min rectangle area | Max rectangle area |
|---------------------|--------------------|
| Small object        | 0 × 0              |
| Medium object       | 32 × 32            |
| Large object        | 96 × 96            |

2. Object detection in Missile-borne Image

2.1. Characteristics of missile-borne image detection

At present, the main bottleneck that restricts the performance of missile-borne image object detection algorithm lies in the overly complicated and blending scene changes. Due to its special application scene, the imaging quality of missile borne camera is closely related to the missile body’s air motion attitude and the battlefield environment. In most cases, the missile-borne image carries the following problems:

a) High real-time requirement

The short flying time of missile in the air and the brief time from capturing object image to landing put forward higher requirements for the speed and efficiency of object detection process.

b) Complex background

In the process of missile motion, the background image captured by missile borne camera is constantly changing, and the changing weather will also affect the imaging quality of missile-borne image. For example, the image clarity will be affected to a certain extent under the conditions of cloudy, rainy, fog and haze, producing a lot of noise in the process of detection and recognition.

c) Relative small object

Combined with the application environment of missile borne camera, it can be seen that when the missile borne camera starts to work at the terminal guidance stage, the far distance to the object and the smaller object lead to the fact that few features can be extracted. In this scene, object detection and recognition algorithm fail to perform well.

2.2. YOLOv3 object detection

Referring to network structures of SSD [4] and ResNet [5], YOLOv3 designs Darknet53, a basic model of classified network. Compared with VGG-16, the common feature extraction network for object detection, Darknet53 reduces computational complexity of the model.

In YOLOv3 network, the structure of Darknet53 is used for image feature extraction, and the yolo structure for multi-scale prediction. In the specific process, a series of convolution operations are carried out by outputting the feature maps (13×13 pixels and 1024 dimension) from the Darknet53. Based on the above sampling, the minimum scale yolo layer is formed after connecting the shallow layer feature maps, and the 13×13×512 channel feature map extracted from layer 79 is convoluted to channel 256, and then the 26×26×256 feature map is generated by sampling. Furthermore, combining it with the feature map of layer 61 and convoluting them to form the mesoscale layer yolo. And the large-scale layer yolo is obtained by corresponding convolution operation of layer 91 and layer36. The feature map prediction, including the coordinate information about the coordinates X and Y of the bounding box, width W and height H, prediction object confidence IOU and category prediction score of the grid. Taking the two buildings striking the object as an example, the number of channels is (4+1+2)×3=21, and the three scale feature maps are 13×13×45, 26×26×45, 52×52×45 respectively, and the structure of Darknet53 can be shown in Figure 1.
After using the data set to directly detect the YOLOv3 network, we find that the examination is difficult under the two conditions: the rotation blur exists and the size of the object is less than dozens of pixels. Therefore, the improved YOLOv3 network mentioned later enhances the detection of small-scale fuzzy objects.

3. Missile-borne Image Object detection Based on the Improved YOLOv3

3.1. The improved network structure

To ensure the running time and the detection effect of small objects with 15 to 30 pixels, we change the convolution module in the original network to multi-branch scale one, making the network more adaptable to different object sizes. It also reduces the network layers to balance the depth and width of the network [6]. Spatial aggregation, based on multi-branch scale convolution, can be completed through low dimensional embedding. Moreover, its ability to feature extraction and representation for image network does not decline. Its internal structure can be divided into three branches, as shown in Figure 1.

![Figure 1. Multi-branch structure](image)

Among them, DepthConcat is to link the feature maps of branch convolution according to the depth to form a map with constant size and depth superposition, which is followed by a 1x1 convolution layer. It does not change the height and width of the map. However, it changes its depth to achieve dimension reduction, facilitating the linear combination of multi-channel features to integrate information among multiple paths. Adding the shortcut layer to overlay the...
features extracted from the previous layer with the current one can avoid gradient dispersion. Only a part of the information can be extracted from each convolution layer in the forward propagation process of the network. The more times they spread forward for small objects, the less information can be learned and retained. Therefore, it is very likely to cause an under-fitting result. Adding the shortcut structure means that adding all the information of the final convolution image in each module equals retaining more information of the initial features to a certain extent. After adding the shortcut structure, the network becomes an optimal output selection model. The output result is optimally learned from the previous block and combination convolution. The network can be regarded as a parallel structure. Finally, the feature extraction combination of different levels can be obtained so that the network can learn the most suitable model and parameters.

Taking the feature pyramid network FPN [7] as a reference, the multi-scale fusion method is used to predict. The feature map of large-scale 52x52 (the first group of multi-branch module output) can provide resolution information for small objects [8]. The last two layers of the multi-branch module simultaneously provide resolution and semantic information for regular and small-scale objects. Different scale detection can effectively detect objects with different scales. Even if the same object is detected in multiple feature layers, the detection effect will not be affected because the best one can be obtained by non-maximum suppression. The specific structure is shown in Figure 2.

3.2. Other improvements

3.2.1. Anchor box calculation

It is not suitable to use the original anchor box to train their small object data, and its setting will affect the accuracy and speed of object detection. We cluster the candidate boxes of a small object dataset in the spatial dimension and calculate the optimal anchors’ number. The anchor’s number and dimension in YOLOv3 are obtained by clustering VOC and COCO datasets [9] [10], which are not suitable for long-distance small object detection. Therefore, its number and the dimension of width and height are redefined.

The k-means clustering algorithm, based on unsupervised learning, is used to cluster and analyze the object frame, extract all the object frame sizes of the data sets, and classify the similar objects into the same category. Finally, we can obtain number k, the number of anchor frames as the scale parameters set during training. K-means clustering analysis uses Euclidean distance, which means that a more extensive scale frame will produce more errors. Therefore, the network loss value uses the IOU overlap degree as the evaluation index. The intersection of the candidate box and the real box is divided by the union to avoid the error caused by the larger BBox compared with the smaller BBox. The function, which replaces the Euclidean distance, is:

\[ d(box, centroid) = 1 - IOU(box, centroid) \]  

The clustering objective function is:

\[ S = \min \sum_{i=0}^{k} [1 - IOU(box, centroid)] \]

As we can see, box represents the candidate box, truth is the object real box, and K is the number of anchor boxes.

3.2.2. Data enhancement

In the process of network iteration, 400 data sets are expanded by adding data enhancement [11]. Besides, the following factors are added: affine transformation, rotation blur, jitter blur, brightness, right and left, up and down, flip, hue, saturation, Gaussian
noise, sharpening, proportion multiplying pixel, piecewise affine, snow, cloud, fog, snow and other climate situations. These data enhancement methods are stored in the sequence. When the image data is actually read, the original image and the object marker frame are enhanced at the same time in a random way, which not only increases the data of simulating a complex environment but does not affect the position of the marker box.

Through data enhancement in network training, the distributed data density is increased synchronously in the spatial dimension. The increased data set has a significant improvement in the ability to object recognition and network generalization. Table 2 shows the comparative experimental results.

| Data enhancement mode          | Number of test sets | Recognition accuracy |
|--------------------------------|---------------------|----------------------|
| No enhancement                 | 1000                | 71.4%                |
| Single data enhancement        | 1000                | 86.7%                |
| Multinomial random superposition | 1000            | 90.3%                |

3.2.3. **Non-Maximum Suppression**

Non-Maximum Suppression (Non-Maximum Suppression, NMS [12], its effect is to suppress the interval part which is not the maximum value, and it can also be called local maximum search.

The improved YOLOv3 network can output lots of object boxes directly. There are about 8 object boxes detected by the image collected by the camera with a resolution of 640×480. We want to eliminate multiple overlapping detection boxes of one object and only reserve the best one for each object. Here, we use the non-maximum suppression method to suppress the elements that are not the maximum to refine the number. Traversing the highly overlapped boundary boxes helps us preserve the prediction box with the highest retention reliability, and non-maximum suppression is performed separately for each class of objects.

The detection effect is shown in Figure 3. The left one shows the detection result after passing through the network, and the right shows the result after suppression.

![Figure 3. Comparison before and after non-maximum suppression](image)

4. **Experiment Simulation**

The operating system of the experimental platform is Ubuntu 16.04, the processor is Intel Xeon E5-2620v4, 2.10GHz, GPU is Nvidia 1080ti.
4.1. Small object detection results

Data enhancement is added to both YOLOv3 and the improved YOLOv3 network experiments, but the latter plays a more significant role in improving small objects’ detection accuracy. Table 3 shows the comparative experimental results. The number of recognized images is 1000, and the object size is between 25 and 100 pixels. The detection effect can be seen in Figure 4. The first column represents the detection effect of the YOLOv3 algorithm, and the improved one produces the second. It indicates that the improved algorithm performs better in detecting the object under such backgrounds as different scales, rainy day, motion blur, rotation blur, and snow day.

Table 3. Comparison of small object detection results between YOLOv3 and improved YOLOv3

| Detection method        | Detected Number | Accuracy |
|-------------------------|-----------------|----------|
| YOLOv3                  | 1000            | 63.7%    |
| Improved YOLOv3         | 1000            | 84.8%    |

Figure 4. Sequence detection results of partial images
From the experiment, we can see that by adding data enhancement and training various noises, the model, to a large extent, makes the detection algorithm more accurate. Using a small amount of data can also make the object detection algorithm carry good robustness and identify small objects in complex background.

5. Conclusion
In this paper, a missile-borne image object detection method based on improved YOLOv3 is proposed. The method is based on a small number of data sets. By adding multi-branch structures and residual shortcut layers, the number of network layers is reduced. Using the K-means clustering method automatically generates anchor area, enhancing the characterization advantage of the feature map and recognizing small objects. It also uses data enhancement and expansion to increase the training’s data diversity, simulating the influence of various factors such as climate, fuzzy, and noise. All these help improve the performance of small object detection in complex backgrounds and boost detection accuracy. As for the object detection research of missile-borne images, this method has specific application value in engineering.

References
[1] B. Singh and L. S. Davis, “An analysis of scale invariance in object detection - snip,” 2018.
[2] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, “Few-shot object detection with attention-rpn and multi-relation detector,” 2020.
[3] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv e-prints, 2018.
[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” 2016.
[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2016.
[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.
[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” 2017.
[8] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” 2018.
[9] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft coco: Common objects in context,” 2015.
[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” International Journal of Computer Vision, vol. 111, no. 1, pp. 98–136, 2015.
[11] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, “Augmentation for small object detection,” in 9th International Conference on Advances in Computing and Information Technology, 2019.
[12] A. Neubeck and L. J. V. Gool, “Efficient non-maximum suppression,” in International Conference on Pattern Recognition, 2006.