Alongshore Ship Detection Based On Multiscale Feature Fusion of Rotation Region Proposal Networks

Hao Dou1,*, Jiaxing Mao2,a, Zhihong Pan2,b
1 The 38th Research Institute Of China Electronics Technology Group Corporation, Hefei, China
2 School of Artificial Intelligence And Automation, HUST, Wuhan, China
*277942607@qq.com, 86-15871707185
a jxingm@163.com, 86-13638688494
b 1628262802@qq.com, 86-13045623334

ABSTRACT In this paper, we propose an alongshore ship detection method based on multiscale feature fusion of rotation region proposal networks, the method provides an end-to-end ship detection framework, which is divided into three modules: multiscale feature module, rotation region proposal network module and context rotation region pooling module. The multiscale feature module integrates multiscale features, the rotation region proposal module can generate rotating bounding boxes in any direction, the context rotating region pooling module fuses the context information to pool the rotating region and complete the classification and position regression of bounding boxes. Experimental results show that our method can achieve good performance in remote sensing data sets.

CCS Concepts: Computing methodologies ~ Artificial intelligence ~ Computer vision ~ Computer vision problems ~ Object detection

1. INTRODUCTION
Semantic segmentation and target detection are important tasks in scene understanding. Semantic segmentation is an image understanding method based on pixel level. Compared with target detection, semantic segmentation can obtain more detailed image information, which can be used to get the target detection result. However, in the process of semantic segmentation, training samples need to be marked at the pixel level, the samples need to be marked at the pixel level, which is complex and consumes more time. In addition, semantic segmentation network needs up-sampling and down-sampling operation of feature maps, which consumes more computing resources. Therefore, compared with semantic segmentation, the annotation of target detection is simple and requires less resource consumption.

At present, some efficient target detection models are proposed, which can be roughly divided into two categories: the first class of models which adopts region generated networks (RPN), such as Fast-RCNN [1], FPN [2], R-FCN [3]; the second class of models do not use region generated networks, such as YOLO [4], SSD [5]. The former model divides target detection into two stages: it first obtains horizontal bounding box by Region Proposal Networks, and then utilizing non-maximum suppression (NMS) for selecting high confidence bounding boxes to classify and regress; the latter model regards the target detection as a regression problem, which directly obtains the category and position information of target according to the feature map. The above models both adopt horizontal bounding
box to mark the target position, which has achieved remarkable results in the natural image, but the horizontal bounding box is not a good way to mark the target [6] in some scenarios. To the ship target in remote sensing image, the length of which is relatively large and its direction is arbitrary, in addition, the position of ship exists parallel and dense distribution. On the one hand, when the ship is inclined and the horizontal bounding box is used, there exists a lot of redundant information in the box. On the other hand, when the ships are densely distributed, adopting the horizontal bounding box is not conducive to distinguish each ship target. At the same time, the large overlap area causes non-maximum suppression to lose target, therefore, these models are not suitable for ship target detection in remote sensing image.

Compared with the horizontal bounding box, Rotation Region Proposal Networks (RRPN) can generate the rotating bounding box inclined in any direction, which can avoid the problems caused by the horizontal bounding box in the process of ship detection. The spatial resolution of low-level feature in CNN is large, which contains more details and structural information, at the same time, the spatial revolution of high-level feature in CNN is small, which contains more semantic Information. Therefore, fusing different level features can express the target fully. Because some areas in the background that are very similar to the structure and texture of the target, only considering the information of targets themselves can make these areas be misclassified. To solve this problem, context information is adopted, which is helpful to make correct classification. we propose an alongshore ship detection method based on multiscale features fusion of rotation region proposal networks, the method provides an end-to-end ship detection framework, which is divided into three modules: multiscale feature module, rotation region proposal network module and context rotation region pooling module. The multiscale feature module integrates multiscale features, the rotation region proposal module can generate rotating bounding boxes in any direction, the context rotating region pooling module fuses the context information to pool the rotating region and complete the classification and position regression of bounding boxes.

The rest of this paper is organized as follows. The second part of this paper introduces the related work; the third part presents the structure and structure of our algorithm in detail; the fourth part presents our experimental results; the fifth part is a summary of this paper.

2. RELATED WORK

For targets with arbitrary direction and dense distribution, such as text in natural scenes and ships in remote sensing images, it is more appropriate to use the rotating bounding box to mark the target than the horizontal bounding box. The rotating bounding box not only needs to obtain the length and height of the target, but also it needs to obtain the rotation angle of the target. In order to obtain the rotating bounding box, these methods based on CNN are roughly divided into two categories: the first one category is based on semantic segmentation [7][8], the second one obtains the rotating bounding box directly through the network [9] [10]. The former generally first extracts the region of the target through the semantic segmentation network, and then it utilizes some post-processing methods to obtain the rotating bounding box. In the processing, the semantic segmentation is utilized, which needs more resource consumption; the latter uses the network to obtain the rotating bounding box directly, which can realize the end-to-end detection process, such as R2CNN model [9] and RRPN model [10]. R2CNN model introduces the information of rotation angle in the stage of target classification and position regression, which can obtain the rotating bounding boxes directly, but the obtained bounding boxes is horizontal in the RPN stage, when the target is distributed densely, Non-maximum suppression may cause the loss of target; RRPN model introduced rotation angle information both in RPN stage and in the stage of classification and position regression, which can obtain better rotating bounding box. However, RRPN model ignored the context background information in the process of region pooling, which leads the expression of the object is not enough accurately.
3. OUR PROPOSED METHOD

The framework of our proposed method is shown in Figure 1, which mainly consists of three modules: multiscale feature module, rotation region proposal network module and context rotation region pooling module. The multiscale feature module integrates multiscale features, the rotation region proposal module can generate rotating bounding boxes in any direction, the context rotating region pooling module fuses the context information to pool the rotating region and complete the classification and position regression of bounding boxes.

3.1 Multiscale features

The low-level feature of CNN can provide more texture and structure information, at the same time, the high-level feature of CNN has large receptive field, which can obtain high-level semantic information. For ship targets, they often have long and thin structure, the length is generally large, and the width is small. Besides, there exists different kinds of ships and they have different sizes. Therefore, we adopt multiscale feature fusion network to adapt to different scale feature space. The structure of multiscale feature network is shown in Figure 2. VGG16[11] is used as the basic network and its full connection layer and the last pooling layer are removed. Here, the last convolution layer of convolution block3, 4 and 5 is named conv3, conv4, and conv5, receptively. As shown in Figure 2, the low-level features are pooled and the high-level features are de-convoluted [12], which can transform different level features to have the same size. At the same time, in order to improve the generalization ability of the model, it is necessary to adjust the characteristic value of different levels to the same range [13], therefore, we adopt L2 normalization to compete this operation. At last, the depth of the feature map is reduced to the same size as conv5 by 1x1 convolution kernel.

\[ \hat{x} = \frac{x}{\|x\|_2} \]

Figure 1 The framework of our proposed model

Figure 2 The multiscale feature maps
\[ \|x\|_2 = \left( \sum_{i=1}^{d} |x_i|^2 \right)^{\frac{1}{2}} \]

In formula, \(x\) and \(\hat{x}\) represent the original pixel vector and the normalized pixel vector, respectively. \(d\) represents the depth of the feature map.

If only a simple normalization is performed on the feature map, the normalization result may be very different from the original value range of the feature map, resulting in a decrease in learning speed \([13]\). Therefore, inspired by BN layer \([14]\), a learnable scaling factor \(\gamma\) is introduced into L2 normalization, which can scale the normalized results and the scaled value is defined as \(y_i\).

\[ y_i = \gamma_i \hat{x}_i \]

According to the chain rule, during the training process, the back propagation gradient of L2 normalized layer is:

\[
\frac{\partial l}{\partial \hat{x}} = \frac{\partial l}{\partial y} \cdot y \\
\frac{\partial l}{\partial x} = \frac{\partial l}{\partial \hat{x}} \left( \frac{1}{\|x\|_2} - \frac{xx^T}{\|x\|_2^2} \right) \\
\frac{\partial l}{\partial y_i} = \sum_{y_i} \frac{\partial l}{\partial \hat{x}_i} \hat{x}_i
\]

In formula, \(y = [y_1, y_2, \ldots, y_d]\).

### 3.2 Rotation Region Proposal Network

The function of the rotation region proposal network is to generate rotating bounding boxes in any direction. The structure of the network is similar to RPN network in Faster-RCNN \([1]\). The difference is that we need to add rotation angle into the expression of bounding box and anchors, also, rotation angle need to be introduced into non maximum suppression. In the following, the expression of rotating bounding boxes and rotation anchors are introduced, also, rotating non maximum suppression and loss function are introduced too.

#### 3.2.1 Rotating bounding box

The representation of horizontal bounding box is relatively simple, which can be directly represented by the coordinates of the upper left corner and the lower right corner \((x_{\min}, y_{\min}, x_{\max}, y_{\max})\). As the rotating bounding box needs to introduce rotation angle, this representation method is no longer suitable. Here, we define rotating bounding box as \((x, y, w, h, \theta)\). As shown in Figure 3, \(x\) and \(y\) represent the center of the rotating bounding box, \(w\) and \(h\) represent the longer side and the shorter side length of the rotating bounding box, respectively. \(\theta\) represents the angle between the long edge of the bounding box and the horizontal direction, the angle range is set to \([-\frac{\pi}{4}, \frac{3\pi}{4})\).
3.2.2 Rotation anchors

In the RPN network of fast RCNN, each pixel on the feature map corresponds to k candidate windows, which are called anchors. The anchors ignore rotation angle and they only consider the aspect ratio and scale information. In order to generate rotating bounding box, rotation anchors (R-anchors) are defined, the rotation angle parameter is introduced into anchors, therefore, R-anchors correspond to three parameters: aspect ratio, scale and rotation angle. According to the structural characteristics of the ship target, the aspect ratio is set to 1:3, 1:5, 1:7 and 1:9, the scale is set to 4, 8, 16 and 32, and the rotation angle adopts 6 angles: $\frac{\pi}{6}$, 0, $\frac{\pi}{3}$, $\frac{\pi}{2}$ and $\frac{2\pi}{3}$. Here, each R-anchor is represented by $(x, y, w, h, \theta)$, therefore, each pixel of feature map in rotation region network corresponds to 96 R-anchor(4×4×6), the depth of the classification layer is 192 (2×96) and the depth of regression layer is 480 (5×96).

3.2.3 Skew Non Maximum Suppression

The number of target bounding box generated by RPN network is large, in order to improve the efficiency of detection, Intersection over Union(IOU) and Non-Maximum Suppression (NMS) are used to select a certain number of bounding boxes with high confidence for classification and location regression. For ship targets, the direct use of Non-Maximum Suppression and intersection over union may lead to the loss of targets, therefore, we need to adopt skew IOU and Skew NMS for rotating bounding box in RRPN.

The calculation method of Skew IOU is shown in algorithm1. As shown in Figure4, firstly, adding intersection point M and N into PSet, secondly, competing sort counterclockwise of vertex in PSet and calculating area of the polygon NHMB by triangulation [15], $S_{NHMB} = S_{\Delta NHM} + S_{\Delta NHB}$. Finally, according to the ratio of skew intersection over union, the Skew NMS removes redundant bounding boxes with large overlapping area.

Algorithm1 the calculation method of skew IOU

**Input:** Rotation Matrix $R_1, R_2, \ldots, R_N$

**Output:** Skew IOU[1, N][1, N]

1. IOU[1, N][1, N], The initial value of the matrix is 0
2. For each pair $R_i, R_j(i \neq j)$ do
3. vertex set PSet is set to empty
4. Add intersection point between $R_i$ and $R_j$ into PSet
5. Add vertex of $R_i$ fall inside $R_j$ into PSet
6. Add vertex of $R_j$ fall inside $R_i$ into PSet
7. Compete sort counterclockwise of vertex in PSet
8. Calculate area of polygonal I composed of PSet vertex set by triangulation, and then the intersection area of two rectangles is:
9. $\text{IoU}(i, j) = \frac{I}{\text{Area}(R_i) + \text{Area}(R_j) - I}$
10. end for
11. Return IoU

3.2.4 Loss Function
In the training process of rotation region proposal network, the decision conditions of positive sample $R$-anchors are as follows: (i) Skew IOU > 0.5 and skew angle minus truth skew angle less than $\frac{\pi}{12}$ or (ii) Skew IOU gets maximum; the decision conditions of negative sample $R$-anchors are as follows: (i) Skew IOU < 0.2 or (ii) Skew IOU > 0.5 and skew angle minus truth skew angle greater than $\frac{\pi}{12}$.

$$L(p, l, v^*, v) = L_{\text{cls}}(p, l) + \lambda L_{\text{reg}}(v^*, v)$$

In the formula, $l$ represents category, $l = 1$ represents target, $l = 0$ represents background, $p = [p_0, p_1]$ represents the probability of belonging to the background and target, $\lambda$ is balance factor, which decides weight of classification task and position regression task in loss function. $v = (v_x, v_y, v_w, v_h, v_\theta)$ represents predicted rotating bounding box, $v = (v'_x, v'_y, v'_w, v'_h, v'_\theta)$ represents truth bounding box, the loss function of bounding box classification and location regression network is defined as:

$$L_{\text{cls}}(p, l) = -\log p_l$$
$$L_{\text{reg}}(v^*, v) = \sum_{i \in \{x, y, w, h, \theta\}} \text{smooth}_1(v^*_i - v_i)$$

$$\text{smooth}_1(x) = \begin{cases} 
0.5x^2 & \text{if } |x| < 1 \\
|x| - 0.5 & \text{otherwise}
\end{cases}$$

In the formula, the calculation method of $v$ and $v^*$ is:

$$v_x = \frac{x - x_a}{w_a}, \quad v_y = \frac{y - y_a}{h_a}$$
$$v_h = \log \frac{h}{h_a}, \quad v_w = \log \frac{w}{w_a}, \quad v_\theta = \theta \ominus \theta_a$$

6. Add vertex of $R_j$ fall inside $R_i$ into PSet
7. Compete sort counterclockwise of vertex in PSet
8. Calculate area of polygonal I composed of PSet vertex set by triangulation, and then the intersection area of two rectangles is:
9. $\text{IoU}(i, j) = \frac{I}{\text{Area}(R_i) + \text{Area}(R_j) - I}$
10. end for
11. Return IoU

Figure 4: Skew Intersection over Union
In the formula, $x$, $x_a$ and $x^\ast$ represent predicted bounding box, anchor and truth bounding box; the meaning of $y$, $h$, $w$ and $\theta$ is same as $x$. Besides, $\Theta b = a - b + k\pi$, is integer, the range of $a \Theta b$ is adjusted to $[\frac{-\pi}{4}, \frac{3\pi}{4})$.

3.3 Context Rotation Region Pooling

The context rotation region pooling module includes two parts: one is the context rotation region pooling, the other is the boundary box classification and regression. The network structure of boundary box classification and regression is the same as Faster-RCNN, which includes two full connection layers and a multi-task branch. For ship targets, if the horizontal region pooling is used, when the target is inclined, there will be a lot of redundant information in the horizontal boundary box, therefore, we adopt the rotating region pooling. Besides, there are some regions in the background have similar structure and texture with the target, if we only consider the characteristics of the target, which can make it is easy to distinguish these regions wrongly. Therefore, we introduce context background information in the process of rotation region pooling. Now, we introduce the two parts in detail.

3.3.1 Rotation Region Pooling

Due to the fixed input size of the full connection layer, it needs to be pooled by Region (ROI Pooling), which can transform different size bounding boxes into feature vectors with the same length, the rotation region pooling is shown as Figure 5, along the long and short sides of the rotating bounding box, the bounding box is divided into sub areas with the number $H_r \times W_r$, and then each sub area is pooled to obtain the result of the rotating region pooling.

![Figure 5 Rotation Region Pooling](image)

3.3.2 Context Information

As shown in Figure 6, the rotating bounding box obtained generated by the rotation region network is extended to obtain the bounding box containing context information. Assuming that the rotation bounding box is $(x_p, y_p, w_p, h_p, \theta_p)$ and the expanded bounding box containing context information is $(x_c, y_c, w_c, h_c, \theta_c)$, the relationship between the two bounding boxes can be expressed as follows:

$$
\begin{align*}
x_c &= x_p, \\
y_c &= y_p, \\
h_c &= \alpha h_p, \\
w_c &= \beta w_p, \\
\theta_c &= \theta_p
\end{align*}
$$

The region of the feature map corresponding to the rotating bounding box and the bounding box containing context information should be pooled respectively, and then connect the two pooled results in parallel to get the features including both the characteristics of the target itself and the context.
information of the target. Finally, using the classification of the bounding box and the location regression network to obtain the category and location of the bounding box.

![Figure 6 Context Information](image)

**4. EXPERIMENTS AND RESULTS**

**4.1 Evaluation Dataset and Implementation Details**

The experimental data is collected from Google Earth, which contains 400 images with a resolution of about 1 m, each image is between 1000-3000 pixels in length and width. We randomly select 100 images as the test set, and the rest is regarded as the training set. We adopt rotating bounding box to mark the ship target, the very small ship is ignored and is regarded as the background. For the training set, we obtain 994 training samples with the size 1000×1000 at 600 pixels step. Also, we randomly flip the training samples in the process of training.

The hardware configuration of the experiment is Intel E5 CPU, NVIDIA 1080ti GPU and the software platform is Ubuntu 16.04. The experiment is carried out with Caffe deep learning framework, adopting fine-tuning on the vgg16 pre-training model and the pattern of end-to-end training. The learning rate are is to 10⁻³ in the first 40000 iterations, it is set to 10⁻⁴ in the next 40000 iterations, and the learning rate is 10⁻⁵ in the final 20000 iterations. Also, weight decay is 5×10⁻⁴, momentum is 0.9. The context information bounding box extension parameter is set to $\alpha = 1.2$, $\beta = 0.4$.

**4.2 Performance Metrics**

In this paper, the detection result is qualitatively evaluated by Accuracy and Recall, which are defined as:

\[
\text{Accuracy} = \frac{\text{Total number of correct detection targets}}{\text{Total number of detection targets}}
\]

\[
\text{Accuracy} = \frac{\text{Total number of correct detection targets}}{\text{Total number of actual targets}}
\]

**4.3 Experiments Results**

To illustrate the advantages of our method, we compare our method with Faster-RCNN [1] and RRPN [16]. From the following comparison results, we can see that our method can achieve better performance than other methods.
Table 1 The results of different models

| Method      | Target Type | The Actual number of ship | The number of correct detection ship | Recall (%) | The number of false detection target (%) | Accuracy (%) |
|-------------|-------------|----------------------------|---------------------------------------|------------|------------------------------------------|--------------|
| Faster-RCNN | Big ship    | 394.0                      | 237.0                                 | 82.99      | 37.0                                     | 91.44        |
|             | Small ship  | 92.0                      | 68.0                                  | 73.91      |                                          |              |
| RPN         | Big ship    | 394.0                      | 352.0                                 | 89.34      | 55.0                                     | 88.64        |
|             | Small ship  | 92.0                      | 77.0                                  | 83.70      |                                          |              |
| Our method  | Big ship    | 394.0                      | 365.0                                 | 92.64      | 22.0                                     | 95.31        |
|             | Small ship  | 92.0                      | 82.0                                  | 89.13      |                                          |              |

5. CONCLUSION
To solve the actual problems in the process of ship detection, we propose a method of inshore ship detection based on multiscale feature fusion of rotation region generation network. We introduce inclination angle in the network to obtain rotating bounding box in any direction, which can avoid problems of horizontal bounding box. Besides, the context information is introduced into the network, which can further improve the performance of detection result. The comparison with these classic models proves the effectiveness of our model.

REFERENCES
[1] Ren S, He K, Girshick R, et al.2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149.
[2] Lin T Y, Dollar P, Girshick R, et al. 2017. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 936-944.

[3] Dai J, Li Y, He K, et al. 2016. R-FCN: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 379-387.

[4] Redmon J, Divvala S, Girshick R, et al. 2016. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779-788.

[5] Liu W, Anguelov D, Erhan D, et al. 2016. SSD: Single shot multibox detector. *European conference on computer vision. Springer*, Cham, 21-37.

[6] Fathi A, Wojna Z, Rathod V, et al. 2017. Semantic instance segmentation via deep metric learning. https://arxiv.org/pdf/1703.10277.pdf.

[7] Yao C, Bai X, Sang N, et al. 2016. Scene text detection via holistic, multi-channel prediction. https://arxiv.org/pdf/1606.09002.pdf.

[8] Zhang Z, Zhang C, Shen W, et al. 2016. Multi-oriented text detection with fully convolutional networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4159-4167.

[9] Jiang Y, Zhu X, Wang X, et al. 2017. R2CNN: rotational region CNN for orientation robust scene text detection. https://arxiv.org/ftp/arxiv/papers/1706/1706.09579.pdf.

[10] Ma J, Shao W, Ye H, et al. 2017. Arbitrary-oriented scene text detection via rotation proposals. https://arxiv.org/pdf/1703.01086.pdf.

[11] Simonyan K, Zisserman. 2014. A. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/pdf/1409.1556.pdf.

[12] No H, Hong S, Han B. 2015. Learning deconvolution network for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*. 1520-1528.

[13] Liu W, Rabinovich A, Berg A C. 2015. Parsenet: Looking wider to see better. https://arxiv.org/pdf/1506.04579.pdf.

[14] Ioffe S, Szegedy C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. https://arxiv.org/pdf/1502.03167.pdf.

[15] D. A. Plaisted and J. Hong. 1987. A heuristic triangulation algorithm. *Journal of Algorithms*, 8(3):405-437.

[16] Ma J, Shao W, Ye H, et al. 2017. Arbitrary-oriented scene text detection via rotation proposals. https://arxiv.org/pdf/1703.01086.pdf.