Recognition and classification of water surface targets based on deep learning

Yanhua Peng¹,², Yipu Yan¹, Biao Feng¹, Xingyu Gao¹,²*

¹. School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, Guilin, Guangxi, China, ².Guangxi Key Laboratory of Manufacturing System and Advanced Manufacturing Technology, Guilin, Guangxi, China
email:gxy1981@guet.edu.cn

Abstract. Aiming at the shortcomings of low recognition rate and low calculation rate of surface unmanned ship in complex time-varying water surface environment, a surface target detection method based on Faster R-CNN is proposed in this paper. Firstly, the water surface image was enhanced by McCann Retinex method to improve the image quality under complex background. Secondly, a water surface target data set was established. Finally, based on Faster R-CNN algorithm, VGG, Resnet and Inception network structures were employed to test and analyze the data set. The results show that the detection method proposed in this paper can effectively complete the identification and classification of six categories of common targets on the water surface, which has significant guiding significance for the autonomous obstacle avoidance and maritime search and rescue of unmanned ships.

1. Introduction
With the rapid development of inland and offshore shipping industry, how the unmanned ships can quickly and effectively identify surface targets provide early warning and real-time monitoring information for water safety evasion, personnel search, and sea state monitoring has become one of the important issues in the fields of intelligent identification, informationization and sensors[1-6]. At present, the traditional surface object detection methods have the following problems: firstly, the water surface environment is complex and changeable, the ship and the water surface building body are mostly based on over-the-horizon imaging, the detection accuracy for the target at a particular close range is low, and it is easy to miss or misdetect[3,4]. Secondly, the traditional machine learning method lacks the pertinence to the region selection, this method mainly uses the direction gradient histogram, the scale invariant characteristic, etc to distinguish the sliding window, which is easy to cause redundancy of sliding window and high time complexity [5,6]. Therefore, it is an important challenge to carry out the research on the surface-oriented target recognition technology for shipping and marine security, which is also a bottleneck technology for unmanned ship from theory to industrialization process[1].

Aiming at the difficulties mentioned above, in order to improve the detection performance of the unmanned ship’s water surface vision system in complex scenes, Li [7] estimated the camera parameters and initialized the scene depth, an optical consistency term was added onto the energy function to simulate the image appearance changes caused by scattering, this method not only can estimate the depth of field, but also improve the visibility of foggy video sequences. However, the proposed method is based on the known atmospheric scattering model with many constraints and low real-time computation, and can not be suitably applied in the water unmanned ship visual system.
Based on the constrained contrast histogram equalization (CLAHE) algorithm, Hitam [8] employed the Euclidean norm to enhance the image contrast and to improve the output quality for the results processed by the image RGB and HSV model respectively. However, this method can introduce more noise in some scenarios. Jobson [9] conducted extensive image processing experiments based on the Retinex theory. On this basis, the reference [10] proposed a fast image defogging method based on the dark channel prior and Retinex theory, which combined the idea of image enhancement and image restoration to improve the image degradation caused by bad weather to some extent. In recent years, with the rapid development of deep learning theory, the field of target detection has stepped into a new stage. Du [11] proposed a ship classification method using Hu moment and ART coefficient to extract the shape characteristic value of the ship target area. The experimental results show that this method can effectively distinguish different types of ship targets in the set scenario, but it does not have the applicability of popularization. In view of this, Leclerc [12] used Inception and ResNet network models to classify ships based on transfer learning instead of using random parameter initialization to train CNN, the classification accuracy of the model was improved to some extent. Reference [13], RPN network was used to realize the automatic generation of potential area of water surface target, and the efficiency of the algorithm was improved by combining convolution feature. The target detection method proposed by Ren Faster R-CNN 2015 combines the possible regions of the separated target with the classification of convolutional neural networks, which makes the target detection process more concise and the detection accuracy greatly is improved [14]. On this basis, Qi [15] used image downscaling and scene narrowing method to construct a hierarchical narrowing network, which significantly improved the calculation speed of Faster R-CNN.

Aiming at the aforementioned problems, the identification and classification methods of water surface objects in complex natural environment based on Faster R-CNN are investigated and analyzed. Firstly, a McCann Retinex algorithm based on color constancy theory is adopted [16]. Secondly, based on the inconsistent size of the surface object, VGG-16, Resnet-50, and Inception-v2 network structure models are compared and analyzed. Then, a small data set of surface objects is fabricated, which mainly includes large ships, lifeboats and surface personnel. Finally, based on the existing data sets, the Faster R-CNN algorithms based on the aforementioned three network models are analyzed experimentally. The data show that Resnet-50 and VGG-16 network models have high detection accuracy for surface objects. Inception-v2 network models are more accurate for target extraction with low resolution in the large field of view.

2. Surface image enhancement

The imaging effect of water surface target is easily disturbed by factors such as water mist, material, light wavelength and background water wave, which makes it difficult to realize the segmentation and recognition of the water surface target [17]. Therefore, a McCann Retinex method is adopted based on spiral strategy [16]. The surface target image is pre-processed to improve the quality of the target image, reduce the interference of environmental noise, improve the signal-to-noise ratio, and make the processed surface image more suitable for subsequent analysis and processing.

As shown in Figure 1, firstly, McCann Retinex algorithm compares two pixels with longer spacing, and then gradually moves to the short spacing to compare pixels according to ratio, product, reset and average four-step operation. In each comparison, in the meanwhile, the comparison direction changes clockwise and the distance between the two pixels is gradually reduced.
Let the input graph size be $M \times N$, where the maximum pixel is $Max$, the output image $r_n(x, y)$ is initialized as a constant $t$, and the starting position $(t, 0)$ of the reference point $(0, 0)$ in one iteration path is $t$, which is determined by the size of the original image:

$$t = 2^{\left\lfloor \log_2 \left(\min(M, N)\right) \right\rfloor - 1}$$

(1)

After completing the above iteration, the following equations can be procured:

$$r_{n+1}(x, y) = \frac{r_n(x, y) + r'_n(x, y)}{2}$$

(2)

$$r'_n(x, y) = \begin{cases} r_n(x, y) + \Delta l & r_n(x, y) + \Delta l \leq Max \\ Max & r_n(x, y) + \Delta l > Max \end{cases}$$

(3)

where $r_n(x, y)$ is the result of the last iteration and $r'_n(x, y)$ is the sum of the brightness difference $\Delta l$ between $r_n(x, y)$ the point and the path. $r_{n+1}(x, y)$ is the output value after $n$ iterations. Let $M$ be the maximum of the map, then the pixel transformation of the final output image is expressed as:

$$P(x, y) = \frac{r_{n+1}(x, y) - (r_{n+1}(x, y))_{\min}}{(r_{n+1}(x, y))_{\max} - (r_{n+1}(x, y))_{\min}} \times M$$

(4)

As shown in Figure 2, the effect of surface imaging enhancement in complex environment is compared.

3. Faster R-CNN target detection principle
Faster R-CNN structure is shown in Figure 3, from which it can be found that Faster R-CNN is mainly
composed of two subnetworks, which are region proposed network (RPN) and Fast R-CNN. RPN is responsible for extracting candidate regions from images that may contain targets. Then, Fast R-CNN further classifies these candidate regions and border regression, and output the final detection results.

![Figure 3. Block diagram of Faster R-CNN](image)

### 3.1. Region Proposal Network

RPN network is the structure used to extract the target area suggestion box in Faster R-CNN network, whose structure is shown in Figure 4, where the red box represents the convolution kernel, and the gray mesh represents each pixel point on the feature map. During the training, an overlap rate (IoU) is used to compare the overlap between the candidate window and the real location of the target. The candidate windows with IoU more than 70% of the target's real position are recorded as positive samples, and the candidate windows with IoU less than 30% of the target's real position are recorded as negative samples. Finally, RPN network is trained with a 1:1 ratio of positive and negative samples.

![Figure 4. RPN network structure](image)

IoU overlap rates are defined as:

$$\text{IoU} = \frac{S_{\text{anchorBox}} \cap S_{\text{groundTruth}}}{S_{\text{anchorBox}} \cup S_{\text{groundTruth}}}$$  \hspace{1cm} (5)

The position regression operation of the boundary frame is carried out according to the output anchors translation and scaling in the regression layer, which is formulated as:
\[
\begin{align*}
t_s & = (x - x_a)w_a, \quad t_y = (y - y_a)/h_a, \\
1_s & = \log(w/w_a), \quad t_y = \log(h/h_a), \\
t'_s & = (x' - x_a)w_a, \quad t'_y = (y' - y_a)/h_a, \\
t'_s & = \log(w'/w_a), \quad t'_y = \log(h'/h_a)
\end{align*}
\]  

(6)  

where \(x\) and \(y\) represent the central coordinates of the prediction box, and \(w\) and \(h\) represent the width and height of the prediction box, respectively. \(x_a, y_a, h_a\) and \(x^*, y^*, w^*, h^*\) represent anchor frame and real frame parameters, respectively.

### 3.2. Fast R-CNN

The candidate regions with different sizes are mapped to a fixed size feature vector Fast R-CNN ROI pool layer after obtaining the candidate regions, and then it is input into a series of fully-connected layers to synthesize the extracted features previously. Finally, two parallel fully-connected layers are used to output the detection results, in which the output of the classification layer is the probability distribution of each frame in the foreground and background classification, and the regression layer outputs the border position parameters. For the training classification and bounding box regression, the loss function at this stage is defined as:

\[
L(p, u, r^*, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{reg}(r^*, v)
\]

(7)  

where \(p\) is the discrete probability distribution used to distinguish the categories of the region of interest, \(u\) is the real number of classes, \(r^*\) is the coordinate of the predicted bounding box, \(v\) is the real bounding box of the category \(u\), \(\lambda\) is the hyper parameter which controls the balance between the two tasks, \(L_{cls}\) is the logarithmic loss of real \(u\), \(L_{reg}\) it is the bounding box regression loss, which can be expressed as:

\[
L_{reg} = \sum_{i} \text{smooth}_{u_i}(r^*_i - v_i)
\]

(8)  

\[
\text{smooth}_{u_i}(x) = \begin{cases} 
0.5x^2, & |x| \leq 1 \\
|x| - 0.5|x| > 1 
\end{cases}
\]

(9)  

### 4. Three common target extraction networks

In this paper, VGG-16 [18], Resnet-50 [19] and Inception-v2 [20] are used to extract image features, the basic structure of CNN includes input layer, convolution layer, pooling layer, full connection layer and output layer. As the network hierarchy deepens, the ability of feature extraction is also increases. From Figure 5(a), it can be seen that VGG-16 network model has 16 weight layers, including 13 convolution layers and 3 fully-connected layers, The network model uses a large number of 3×3 convolution kernels in combination and stack to replace the original large convolution kernel, which can effectively reduce the number of network parameters and the computational complexity. Resnet-50 is based on the existing training depth network. Based on the existing training depth network, Resnet-50 is a residual learning framework with the advantages of easy optimization and small computing burden. By adding its own mapping layer, resnet-50 can keep the network from disappearing when the number of layers is very deep, so as to obtain effective training. Resnet-50 contains 49 convolutional layers and 1 fully-connected class, as shown in Figure 5(b), compared with the VGG-16, the number of network layers is deeper. However, due to the use of global average pool operation instead of full connection dense layer, the size of resnet-50 model is more concise. Inception-v2 network increases the network’s scaling adaptability by increasing its width, using convolutional core connections of different sizes to enrich each layer of information in each Inception module. Later, the BN algorithm is used to accelerate the convergence speed of the network, and the existing dense connection layer is used to approximate a sparse structure, avoiding the overfitting problem caused by parameter increase, and improving the accuracy of the model, whose structure is shown in Figure 5(c).
5. Experimental results and analysis

5.1. The surface object data set is established

Because the existing data sets are concentrated on common objects and are mainly labeled objects, they cannot be used for the training of surface objects. In order to realize the recognition and classification of target objects in complex scenes, the source of this data set is mainly divided into two parts: (1) real-time shooting of water surface. (2) data collection of Internet crawler. Water surface real-time shooting is the main source, Internet crawler data is an auxiliary source.

So far, 1189 images of surface targets have been collected. In order to avoid overfitting existing data during training and to enhance the generalization ability of the model, rotation, clipping, noise, flipping, contrast enhancement and other methods are adopted to enhance the diversity of data in this paper. The enhanced dataset contains 3461 images, which includes 988 large ships, 602 lifeboats, 303 reefs, 497 beacon lights, 251 surface personnel and 820 rig. Some of the samples are shown in Figure 6.

Figure 5. Schematic diagram of network structure

Figure 6. Examples of training and test samples
The next step is to use the LabelImg tagging tool to classify and label the collected water surface target images. The tagging information format is the same as that of the PASCAL VOC data set. The specific format is that five parameters are used to make a data set, including category sequence (index), x, y direction coordinates, and the width and height of the target center. As shown in formula 10, where \( x_{\text{max}} \) and \( y_{\text{max}} \) are the lower right corner of the coordinates, \( x_{\text{min}} \) and \( y_{\text{min}} \) are the upper left corner of the border coordinates, and \( w \) and \( h \) are the image width and height.

\[
\begin{align*}
    x &= \frac{x_{\text{max}} + x_{\text{min}}}{2w}, \quad y = \frac{y_{\text{max}} + y_{\text{min}}}{2h} \\
    y &= \frac{y_{\text{max}} - y_{\text{min}}}{h}, \quad h = \frac{y_{\text{max}} - y_{\text{min}}}{h}
\end{align*}
\]  

(10)

5.2. Experimental environment and data analysis

The ship data set used in this article contains 3461 images, 2942 pictures are randomly selected as the training set and 519 pictures as the test set. The experiment was based on the Ubuntu 16.04 operating system, with Pytorch as the learning framework, Intel Xeno Scalable 5120-CPU @2.50Ghz, 256G memory, Cuda9.0, and NVIDIA Quadro GV100 32G graphics card. During the experiment, GPU was called for acceleration operation, and a total of 100 iterations were carried out. The initial number of iterations was 30, the learning rate was set at 0.001, and the learning rate for the next 70 iterations was set at 0.0001.

In order to further evaluate the performance of various algorithms, and considering the large difference in the number of samples of each category, this paper adopts the category weighted average accuracy (wAP) standard to measure the model accuracy. Let the weights \( \alpha_i \) of each category be proportional to its sample size \( M_i \), and there are \( N \) categories, wAP can be expressed as:

\[
wAP = \frac{\sum_{i=1}^{N} |AP| \times \alpha_i}{\sum_{i=1}^{N} \alpha_i} = \frac{M_i}{M}
\]  

(11)

When the overlap between the predicted bounding box and the correct bounding box is more than 50%, the predicted bounding box is deemed to be correct, and the detection results on the whole test set are shown in Table 1.

Table 1. Comparison of target detection results using different network models

| Detection algorithm       | AP (%)       | wAP (%) |
|---------------------------|--------------|---------|
| Ship                      | Lifeboat     | Rig     | Beaconlight | Reef | People |         |
| Faster R-CNN+VGG-16       | 89.32        | 90.80   | 98.75       | 95.51 | 91.68   | 79.01   | 91.97   |
| Faster R-CNN+Resnet-50    | 95.16        | 92.94   | 99.60       | 97.78 | 97.45   | 86.69   | 95.63   |
| Faster R-CNN+Inception-v2 | 81.56        | 83.42   | 88.05       | 87.02 | 85.21   | 65.74   | 83.08   |

From the aforementioned comparative analysis, it can be found that the Resnet-50 model with rich structural features and considerable depth features has better detection capability when facing difficult targets under complex background such as surface personnel and lifeboats. The Resnet-50 network has similar detection capabilities to the VGG-16 network, and is superior to the Inception-v2 network when it comes to targets with complex structures such as drilling platforms and ships. However, the detection accuracy of such targets as surface personnel is limited by small sample size and low resolution, thus resulting in low target identification accuracy of personnel. In addition, in order to more intuitively feel the detection difference between different algorithms, some detection images are selected for comparative analysis, among which, Figure 7(a) is the detection effect diagram of Faster R-CNN+VGG-16 algorithm, Figure 7(b) is the detection effect diagram of Faster R-CNN+Resnet-50, and Figure 7(c) is the detection effect diagram of Faster R-CNN+Inception-v2.
6. Conclusions

In this paper, the surface object imaging is enhanced based on the McCann Retinex method to improve the image quality. Secondly, a small surface object data set containing 3461 surface images is proposed, most of which are large ocean-going ships, drilling platforms and navigation lights. Finally, the performance of three target detection network models based on the Faster R-CNN algorithm in the data set is analyzed. Looking at the data, the Faster R-CNN+Resnet-50 has a higher accuracy than the Faster R-CNN+VGG-16 and the Faster R-CNN+Inception-v2. It can be seen from the sample of the detection results that the three types of Faster R-CNN algorithm can make a detailed prediction for the target, but the accuracy is affected because some relatively rough artificial labels in the data set are inconsistent with the real information. In addition, by analyzing the samples with complex background, it can be found that the models with rich structural feature information and relatively deep feature information have better processing and generalization ability when facing difficult targets. So far, the target data set is still in an early stage, we need not only expand the amount of data need to increase the diversity of data, and we found that the different resolution of the image characteristics influence the performance of detection, especially form different surface target size, so the next step will be aimed at the influence of different resolution image characteristics on detection performance analysis, to enhance the surface under the over-the-horizon target detection performance.

Acknowledgments

The authors gratefully acknowledge the financial support of the NSF of China (No. 51805104), Guangxi Science and Technology Base and Talent Special Program (Gui Ke AD18281038), Guangxi Key Laboratory of Manufacturing System and Advanced Manufacturing Technology Director Project Fund (1725905008Z) and Graduate Innovation Project of GUET (2020YCXS005).

References

[1] Li, Q., Mou, L., Liu, Q., Wang, Y., & Zhu, X. X. (2018). HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 56(12), 7147-7161.
[2] Geng, X., Wang, Y., Wang, P., & Zhang, B. (2019). Motion plan of maritime autonomous surface ships by dynamic programming for collision avoidance and speed optimization. Sensors, 19(2), 434.
[3] Li Yang. (2020). Research on Surface Target Recognition and Tracking based on unmanned ship vision (Master thesis, Dalian Maritime University).
[4] Zheng, J., Su, T., Zhang, L., Zhu, W., & Liu, Q. H. (2014). ISAR imaging of targets with
complex motion based on the chirp rate–quadratic chirp rate distribution. IEEE Transactions on Geoscience and Remote Sensing, 52(11), 7276-7289.

[5] Zhao, H., Wang, Q., Huang, J., Wu, W., & Yuan, N. (2014). Method for inshore ship detection based on feature recognition and adaptive background window. Journal of Applied Remote Sensing, 8(1), 083608.

[6] Zhai, L., Li, Y., & Su, Y. (2016). Inshore ship detection via saliency and context information in high-resolution SAR images. IEEE Geoscience and Remote Sensing Letters, 13(12), 1870-1874.

[7] Li, Z., Tan, P., Tan, R. T., Zou, D., Zhiying Zhou, S., & Cheong, L. F. (2015). Simultaneous video defogging and stereo reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston. pp. 4988-4997.

[8] Hitam, M. S., Awalludin, E. A., Yussof, W. N. J. H. W., & Bachok, Z. (2013, January). Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In: 2013 International Conference on Computer Applications Technology. Zurich. pp. 1-5.

[9] Jobson, D. J., Rahman, Z. U., & Woodell, G. A. (1997). Properties and performance of a center/surround retinex. IEEE Transactions on Image Processing, 6(3), 451-462.

[10] Liu Haibo, Yang Jie, Wu Zhengping, Zhang Qingnian & Deng Yong.(2015).Fast single image defogging method based on dark Channel Prior and Retinex theory. Acta Automata Sinica (07),1264-1273.

[11] Du, Q., Zhang, Y., Yang, X., & Liu, W. (2017, August). Ship target classification based on Hu invariant moments and ART for maritime video surveillance. In: 2017 4th International Conference on Transportation Information and Safety. Banff. pp. 414-419.

[12] Leclerc, M., Tharmarasa, R., Florea, M. C., Boury-Brisset, A. C., Kirubarajan, T., & Duclos-Hindé, N. (2018, July). Ship classification using deep learning techniques for maritime target tracking. In: 2018 21st International Conference on Information Fusion. Cambridge. pp. 737-744.

[13] Li Chang. (2016). Research on Surface Target Detection and Recognition Algorithm (Master's thesis, Huazhong University of Science and Technology).

[14] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149.

[15] Qi Liang, Li Bangyu & Chen Liankai.(2020). Vessel target detection algorithm based on improved CFCN. China Shipbuilding (S1), 40-51.

[16] Funt, B., Ciurea, F., & Mccann, J. (2004). Retinex in matlab™. Journal of Electronic Imaging, 13(1), 112-121.

[17] Zeng Wenjing.(2013). Research on Surface Target Detection and Tracking of Unmanned craft based on Light vision (doctoral dissertation, Harbin Engineering University).

[18] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[19] 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. Las Vegas. pp. 770-778.

[20] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.