Marine Object Recognition Based on Deep Learning

Bo Shi[1], Hao Zhou[2,*]

[1]School of Automation Huazhong University of Science and Technology, Wuhan, China

[2]School of Transportation School of Automation Wuhan University of Technology Wuhan, China

[*]zhzmq@whut.edu.cn

Abstract. In the research of unmanned surface vessel (USV), accurately perceiving the environment around the USV and recognizing the obstacles in real time are the major difficulties. The existing methods based on lidar or unmanned air vehicle have got good performance, but time and money costs are not what we can afford. After analyzing the difficulties existed in the obstacle avoidance test for USV, we propose a new method called marine object detection based on Single Shot MultiBox Detector (SSD). It solves these difficulties well, and the time and money costs are acceptable to us. After modifying and optimizing the SSD model, its average precision is 93.5% and its time cost is 45ms per image (1280*760), which means that it has much better performance than any existing method. The experimental results show that the method can detect object in real time and have great precision, which ensures the safety of USV during the navigation.

Keywords: Unmanned surface vessel, deep learning, Single Shot MultiBox Detector, Marine object detection.

1. Introduction

The concept of USV has been proposed for a long time and USV has been used to conduct varies scientific research, including marine surveying [1], military and commercial applications[. Until now, there have been many exciting research results. Some companies have developed a variety of USVs, such as mapping, monitoring and fire-fighting USVs[2][3]. But these USVs have relatively small size, the length is under 4 meters and displacement is about two tons. With the development of the national marine industry and the national emphasis on the ocean, the demand for multi-functional large-scale USV has become more and more intense. Large-scale USVs have many great advantages in transport and they can also play an important role in national defense[5]. Therefore, most countries in the world are accelerating the research on large-scale USV. In the research, it is recognized that perceiving the environment around the USV accurately and recognizing obstacles in real time are the most troublesome and important. And the safety of USV during the navigation could not be ensured if it could not recognize obstacles accurately and in real time. The large-scale USVs' length is over 50 meters and the displacement is beyond 100 tons, which shows that it has great military and economic value, so it is necessary to accurately perceiving the environment and recognizing the obstacles to ensure the safety of USV during the navigation. And this paper focuses on recognizing the obstacles accurately and in real time.

Until now, there are many research progresses. The method using remote sensing images to recognize the obstacles has got good performance, but we could not bear the delay and money cost[8][9]. Another method using an unmanned air vehicle (UAV) to follow the USV and transfer the images back to USVs to recognize the obstacles. Obviously, it needs the UAV follow USV all the time and all the way. However, the UAV's load and endurance are very limited. And the method based on lidar refers to the unpiloted, it recognizes the obstacles by raster map collected from the lidar. But the lidar's effective range is only 200 meters, so it may be only used on small size USVs. Furthermore, the approach combineing pulse radar and lidar sensors gets better performance[11]. It gets relative information between a USV and nearby vessels. Because of the effective range of lidar, its performance is very sensitive to distance. When it is out the effective range of lidar, this method is

invalid basically. In summary, there is currently no suitable way to accurately detect obstacles for large-scale USVs.

In this work, after analyzing the problems of the above methods and the specific application environment of USVs, we decide to solve the problem of recognizing the obstacles by deep learning. Currently, the photoelectric detection equipment is standard on large-scale USVs. So we get image from this device and use our modified SSD model, a deep learning model, to recognize the obstacles accurately and in real time. This paper proposes a method for marine object recognition using ship-borne photoelectric detection equipment after analyzing the problems of the above methods and the specific application environment of USVs. The photoelectric equipment used to get images is essentially a high-definition camera with an effective range of up to 3 kms. The proposed method uses modified SSD model, a deep learning model, to recognize the obstacles by the images collected from the photoelectric detection equipment. Firstly, we modify the Single Shot MultiBox Detector model to make it have less parameters but retain the good performance. Then we collect about 6000 images as the dataset and label them according to the VOC format. Next, we split the into three parts, 80% of dataset is for training, 10% is for validation and 10% is for testing. We train our modified SSD model on GPU 1080Ti and the average precision can reach 95% on validation and testing image. Lastly, we test our model during the navigation of the testing USV on lake. The results indicate that its average precision reaches and stabilizes at 93.5% and it costs about 45ms to detect an image (1280*760). And the performance far exceeds the requirements of the large-scale USVs and could strongly ensure the safety of USV during the navigation.

2. Related Work

There are many similarities between USV and unpiloted, and their common major problem lies in accurately perceiving the surrounding environment and recognizing obstacles in real time. After several years of pilotless automobile development, the solution based on lidar has basically solved the urgent problem in unpiloted. So there are many programs that refer to unpiloted to solve the problems in USVSs.

For USV, the method using the lidar to recognize obstacles learns from unpiloted[6][7]. It has three major steps for obstacle detection and recognition, target extraction, target features and classification method. And the most important part in the method is the extraction of individual target. The specific steps: 1) Receiving Point Cloud Library (PCL) of the lidar; 2) Filtering the PCL and extracting single targets from PCL; 3) Combining the geometric characteristics of the target and k-Nearest Neighbor algorithm to achieve target detection. However, effective detection range of the lidar is within 200 meters. So limited by the effective detection range, this method may be only used on small size USVs. Furthermore, the approach fusing a pulse radar and a lidar has better performance[11]. The relative bearing and range information between a USV and nearby vessels is obtained using radar and lidar sensors, and their motion including the position, heading, and speed is estimated based on a dual filter structure using an extended Kalman filter. First, target objects are detected using both pulse radar and a 3D lidar by applying automatic feature extraction algorithms on their measurements. Second, the relative bearing and range information of the targets is obtained from the extracted features. Finally, the target motion is analyzed in a dual filter structure using an extended Kalman filter. So it has great performance in the effective detection range of the lidar. But beyond the effective range, this approach depends on the pulse radar only. However, the echo data of the pulse radar is sensitive to noise, so it is invalid basically. All in all, its robustness is very poor.

In addition, the scheme using remote sensing images (RS) to recognize target is effective[8][9]. With the development of satellite technology, the resolution of remote sensing image (satellite camera) has reached 0.05 meters which is at an amazing level. And this method has great performance. However, time and money cost is unbearable to us. And another method uses a UAV (unmanned air vehicle) to follow USV, then transfers the images back to USV to recognize target[10]. It is clear that UAV have to follow the USV all the time and all the way. However, the UAV's load and endurance

are very limited. So it is impossible to capture HD images to recognize the obstacles accurately and in real time.

3. Single Shot MultiBox Detector Model

The USV needs to recognize obstacles during the navigation and respond in time according to the obstacle information. Therefore, taking into account real-time requirement, it is necessary to improve the original model to meet these demands.

In the object detection filed, most methods are based on convolutional neural networks[21] and have got high accuracy. After years of development, the convolutional neural network has formed a series of classic networks, such as Lenet[12], Alexnet[13], VGG[14], GooLeNet [15], ResNet [16] and so on. Currently, the models used commonly and having good performance can be mainly divided into two types. One is based on the region proposal, including R-CNN [17][24] series(R-CNN, SPP-net[25], Fast-RCNN[22], Faster-RCNN[23]). And another is based on end-to-end, without the region proposal, including YOLO series (YOLO[18], YOLOv2[19]) and SSD[20]. These models' performance is shown in the following Table 1.

Table 1 Comparing the performance of several models

Model Name	Bcakbone	Fps	VOC2007
R-CNN	VGG	<1	0.48~0.66
SPP-net	ZF-5	<5	0.63~0.82
Fast R-CNN	VGG	<5	0.66~0.70
Faster R-CNN	VGG	5	0.73~0.85
YOLO	/	45	0.63
YOLOV2	/	40	0.78
SSD	VGG	20	0.80

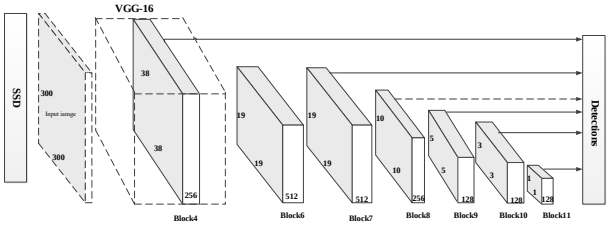

Fig. 1 The network of SSD model based on VGG-16

It is clear in the table, the accuracy of the RCNN series based on the region proposal is higher, and the accuracy of Faster-RCNN model reaches 85%, but the model costs 200ms to detect a picture, which means it cannot meet the real-time requirements. The accuracy of YOLO series based on end-to-end is a little lower, but it also reaches 78% and its detection speed has been greatly improved to 40fps. While the accuracy of SSD is 80%, the detection speed reaches 20fps. Considering the speed and accuracy, the SSD model is used as the object detection model in this paper.

Actually, the object detection task can be divided into two sub-tasks. The first is to classify, and then to locate (regression). The SSD model combines these two tasks, using a modified network based on VGG16 but removing the full connection layer. The network structure is shown in the Fig. 1. And the parameters of the network are shown in the Table 2.

Table 2 The parameters of original SSD model

NO.	Name	Parameters	Times	Note
1	Conv	3*3,64	2	Block1
2	Pool	2*2,maxpool	1	/
3	Conv	3*3,128	2	Block2
4	pool	2*2,128	1	/
5	Conv	3*3,256	2	Block3
6	pool	2*2,maxpool	1	/
7	Conv	3*3,512	2	Block4
8	pool	2*2,maxpool	1	/
9	Conv	3*3,512	2	Block5
10	pool	2*2,maxpool	1	/
11	Conv	3*3,1024	2	Block6
12	Dropout	/	/	/
13	Conv	3*3,1024	2	Block7
14	Dropout	/	/	/
15	Conv	3*3,256/512	2	Block8
16	Conv	3*3,128/256	2	Block9
17	conv	3*3,128/256	2	Block10
18	Conv	3*3,128/256	2	Block11

Obviously, compared with the VGG network structure, the SSD model removes the full-connection layer, which greatly reduces the amount of parameters (the large amount parameters of VGG is mainly caused by last full-connection layer). And the model using multiple continuous small-scale convolution processing instead of large-scale convolution processing makes the network performance better. As for marine object detection, we make the convolution layers (from Block4 to Block11) replaced by inception module[26].

Obviously, compared with the VGG network structure, the SSD model removes the full-connection layer, which greatly reduces the amount of parameters (the large amount parameters of VGG is mainly caused by last full-connection layer). And the model using multiple continuous small-scale convolution processing instead of large-scale convolution processing makes the network performance better. As for marine object detection, we make the convolution layers (from Block4 to Block11) replaced by inception module[26].

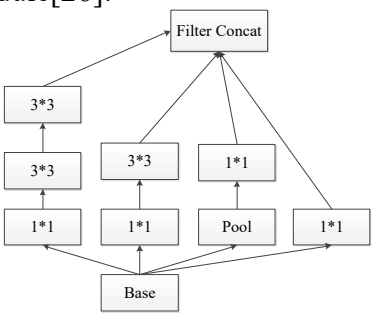

Fig. 2 The network of inception

The inception has three parts. The first part is followed by 1*1 convolution kernel, 3*1 convolution and 3*1 convolution. The second is followed by 3*3 average pooling and 1*1 convolution. The last is 1*1 convolution. Additionally, the inception do not change the size of input feature map. As for the

Block4, the depth of input is 256, so we set the depth of inception module to 128, 96 and 32. The total depth is also 256, but the inception module reduces the parameter amount and improves the performance.

Table 3 The specific parameters of the six layer The size represents the output of the layer. The parameters Size of box and Controller of box decide the set box size for each pixel

Note	Size	Boxes	Size of Box	Box Parameters
Block4	38*38	4	(21,45)	(2,1/2)
Block7	19*19	6	(45,99)	(2,1/2,3,1/3)
Block8	10*10	6	(99,153)	(2,1/2,3,1/3)
Block9	5*5	6	(153,207)	(2,1/2,3,1/3)
Block10	3*3	4	(207,261)	(2,1/2)
Block11	1*1	4	(261,315)	(2,1/2)

In Fig. 2, the SSD model is divided into eleven blocks, and six blocks of them are used as the input of image classification and target location, which shows multiple scales of information for classification and location. Multiple scales make the network more adaptive and it could detect small object on shallow layers and large targets on deep layers.

According to the SSD model, the output of the block4, block7, block8, block9, block10, block11 layers are used to classify and locate. In order to classify and detect on six scales, we put fixed size box on every pixel in feature maps, and the number of box is 4 or 6. So every pixel has k*(4+4+2) outputs. It indicates that every pixel has k boxes, and every box has (4+4+2) outputs, including fixed box location, offset and confidence of every box. It is clear that the categoriesis 2 in this network and the location of actual object is the difference of fixed box location and offset. And the specific parameters of six blocks are shown in the Table 3.

For the layer of block4, the output size of feature map is 38*38, and the number of box for each pixel is 4, so the box sizes are as follows.

$$h_0 = \frac{21}{300},\ w_0 = \frac{21}{300},\ h_1 = \frac{\sqrt{21*45}}{300},\ w_1 = \frac{\sqrt{21*45}}{300} \tag{1}$$

$$h_2 = \frac{21}{300*\sqrt{2}},\ w_2 = \frac{21*\sqrt{2}}{300};\ h_3 = \frac{21*\sqrt{2}}{300},\ w_3 = \frac{21}{300*\sqrt{2}} \tag{2}$$

And the most significant part is loss function, divided into two parts, including classification loss and location loss.

$$L(x,c,l,g) = \frac{1}{N} \cdot L_{conf}(x,c) + \alpha \cdot L_{loc}(x,l,g) \tag{3}$$

$$L_{loc}(x,l,g) = \sum_{i \in Pos}^{N} \sum_{m \in (cx,cy,w,h)} x_{ij}^{k} \cdot smooth_{L1}\left(l_i^m - \hat{g}_j^m\right) \tag{4}$$

$$L_{conf}(x,c) = -\sum_{i \in Pos}^{N} x_{ij}^{P} \cdot \log\left(\hat{c}_i^{\,p}\right) - \sum_{i \in Neg} \log\left(\hat{c}_i^{\,0}\right) \tag{5}$$

$$\hat{c}_i^{\,p} = \frac{\exp c_i^p}{\sum_p \exp c_i^p} \tag{6}$$

The part $L_{conf}(x,c)$is classification loss, x_{ij}^p indicates whether the prediction box i and the real box j match with category p. If they match, they are 1, otherwise, they are 0. $\log c_i^p$ indicates the probability of predicting a match for Category，$\log c_i^0$ indicates predicted probability that the predictive box has no object. The another part $L_{loc}(x,l,g)$ is location loss, x_{ij}^k indicates whether prediction box i and real box j match for category k. If true, it is 1, otherwise it is 0. l_i^m indicates the prediction location, $\hat{g}_j^m$ indicates the real location.

4. Experiment

We train the modified model on our dataset collected from our testing USV and test the model during the navigation of USV on lake.

4.1 Dataset Generation

The main parameters and photoelectric detection equipment of the USV used in this test are shown in Fig. 3.

Fig. 3 The first is the USV used in the test, the total length is about 7.5 meters and the total displacement is 3.5 tons. The second is the photoelectric detection equipment, and the horizontal deflection angle is 360 degrees and the vertical is 60 degrees

Considering that when USVs navigate in the ocean, the surrounding environment is relatively empty. In order to simplify the model, we set two categories of objects, ship and hill. And it is also in line with the actual situation. Based on this assumption, the obstacle objects contained in the training images are mainly ships and hills. During the actual navigation, most of the obstacles are ships. Therefore, we have collected images of various environments and situations. The training images used is about 6000 images totally, including about 1500 images of hill obstacle and 4500 images of ship obstacle. This distribution of data makes training images more in line with reality and may contribute to improve accuracy.

Fig. 4 The images in the dataset the first is ship-obstacle the second is hill-obstacle

After making the dataset, the pictures need to be annotated. The purpose of the object detection is to predict the location of the target in the picture, so the information marked in the picture is the location and object category for the training dataset. In this paper, only two categories need to be detected, hill and ship. The location of the object is saved in the format of (x, y, w, h). In this format, x and y represent the coordinates of the upper left vertex of the rectangle， and w and h represent the length and width of the rectangle containing the target. So the annotation format is (label, x, y, w, h).

4.2 Training

Next, we can train our modified model. We divide the dataset into three parts, 80% is for training, 10\% is for validation and 10% is for testing. In order to accelerate the training speed, GPU 1080Ti is used. During the training, the batch size is 16 for 500 epochs. The inintial learning rate is set to 0.0001, and is divided by 10 at 50% and 75% of the total number of training epochs. Considering that our dataset have few images, the initial weights are pretrained model weights based on COCO and we just do fine-tuning.

4.3 Testing

Table 4 Comparison of some Target Recognition Algorithms

No.	Brief	Reference	MAP	Fps
1	Improved SSD	/	93.5%	>20
2	Based on BP	Chang, X. [9]	95%	<1
3	Based on radar	Ying, S., et al. [27]	81%	<1
4	Based on radar	Alexandrov, C., et al.[28]	84%	<1

After a few hours of training, the final average precision reaches 95% and stabilizes at 95% on validation and testing images. In order to further verify the performance of the model, we test on a testing USV. And the detection accuracy of the object in the actual scene reached 93.5%. Comparing with others, the results are shown in Table 4. Obviously, the method based on UAV and BP network get the best accuracy but the time cost is much higher. Considering the speed and average precision, the SSD model is the most cost-effective.

Fig. 5 The testing result

The Fig. 5 shows the actual test image collected from the photoelectric investigation equipment on the test USV. The ship in the figure can be detected and the location prediction achieves a high accuracy. At the same time, the time cost of each picture (1280*760) is about 45ms which can meet the real-time requirements. Also, the model still need to be improved since there are many problems in the actual detection, like the effectiveness in detecting the small object which is shown in Fig. 6.

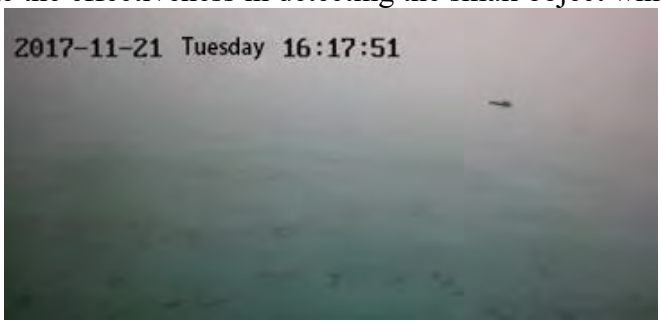

Fig. 6 The small object cannot be detected in the model

5. Conclusion

In this paper, we improve the original SSD model first, and then collect about 6000 images to train the model. At last, the improved model gets better performance, whether in accuracy or real-time. However, this method also has certain flaws. The photoelectric equipment is affected by the weather when collecting images. When it is in heavy fog, rain, snow and other low visibility weather, the model is powerless. In addition, it is unable to detect objects with small size. So it still requires a lot of work to achieve marine object detection in various environments, especially in harsh. And the dataset used to train the model is inadequate, just about 6000 images. If we could get more images and better images, the model would get much better performance.

References

[1]. Wang, J., Ren, F., Li, Z., Liu, Z., Zheng, X., Yang, Y.: Unmanned surface vessel for monitoring and recovering of spilled oil on water. In OCEANS 2016-Shanghai, IEEE, pp.1–4.

[2]. Li, J., Li, T., Li, Y.: NN-based adaptive dynamic surface control for a class of nonlinear systems with input saturation. In Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference. IEEE, pp.570–575.

[3]. Sabanovic, A.: Variable structure systems with sliding modes in motion control—a survey. IEEE Transactions on Industrial Informatics, 2011, 7(2), pp. 212–223.

[4]. Collins, G., Clausse, A., Twining, D.: Enabling technologies for autonomous offshore inspections by heterogeneous unmanned teams. In OCEANS 2017-Aberdeen, IEEE, pp.1–5.

[5]. Campbell, S., Naeem, W., Irwin, G. W.: A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres. Annual Reviews in Control, 2012, 36(2), pp.267–283.

[6]. Wang, S., Xie, L., Ma, F., et~al.: Research of obstacle recognition method for USV based on laser radar, International Conference on Transportation Information and Safety. 2017, pp.343–348.

[7]. Blaich, M., Kohler, S., Schuster, M.,et~al.: Mission integrated collision avoidance for USVs using laser range finder. Oceans. 2015, pp.1–6.

[8]. Zhang, F.: Object recognition and GPU parallel acceleration of optical remote sensing images at sea. Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, 2017.

[9]. Chang, X.: Design and application of marine target recognition system based on UAV infrared remote sensing image. Northeastern University. Liaoning, China, 2013.

[10]. Chen, P.: Recognition Technology Research on Sea Ship Target of UAV Remote Sensing Images. JiMei University. Fujian, China, 2012.

[11]. Han, J., Kim, J., Son, N. S.: Persistent automatic tracking of multiple surface vessels by fusing radar and lidar. In OCEANS 2017-Aberdeen, IEEE, pp.1–5.

[12]. Lecun, Y., Bottou, L., Bengio, Y.: Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998, 86(11), pp.2278–2324.

[13]. Krizhevsky, A., Sutskever, I., Geoffrey, H.: ImageNet Classification with Deep Convolutional Neural Networks.2012, pp.1106–1114.

[14]. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint. 2014.

[15]. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp.1–9.

[16]. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp.770–778.

[17]. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp.580–587.

[18]. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp.779–788.

[19]. Redmon, J., Farhadi.: YOLO9000: Better, Faster, Stronger. arXiv preprint. 2017.

[20]. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., and Reed, S.: SSD: single shot multibox detector. In European conference on computer vision. Springer, Cham, 2016, pp.21–37.

[21]. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 2012, pp.1097–1105.

[22]. R. B. Girshick.: Fast R-CNN. arXiv preprint. 2015.

[23]. Ren, S., He, K., Girshick, R., Sun, S.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Advances in neural information processing systems. 2015, pp.91–99.

[24]. Girshick, R., Donahue, J., Darrell, T., et al.: Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016, 38(1), pp.142–158.

[25]. He, K., Zhang, X., Ren S., Sun, J.: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9), pp.1904–1916.

[26]. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp.2818–2826.

[27]. Ying, S., Chen, J., Shi, C.: Target recognition for marine search and rescue radar. In Natural Computation (ICNC), 2010 Sixth International Conference on Vol. 2. Pp.676–679.

[28]. Alexandrov, C., Draganov, A., Kolev, N.: An application of automatic target recognition in marine navigation. In Radar Conference, 1995. Record of the IEEE 1995 International. Pp.250-255.