Motion Behavior Recognition of Underwater Vehicle Based on YOLOv3

Leisheng Tang1, 2, Jing Hou1*, Hongli Xu3, Mengyan Wu3, Xinyang Xia4 and Yifan Cui5

1Shenyang Jianzhu University, Liaoning, China
2Shenyang Institute of Automation Chinese Academy of Science, Liaoning, China
3Northeastern University, Liaoning, China
4Shenyang University of Technology, Liaoning, China
5Hai Nan University, Hai Kou, China

*Corresponding author e-mail: xxhoujing@sjzu.edu.cn

Abstract. As an important tool for human exploration and understanding of the ocean, underwater vehicle cannot achieve real-time cooperation due to the limitations of underwater acoustic communication. In order to realize the recognition and perception of the cooperative object behavior of underwater vehicle by vision, a new method is provided for the cooperation between underwater vehicle. Select YOLOv3 target detection algorithm, this paper to test the underwater vehicle motion behavior recognition, first of all, the collected experimental data set, using LabelImage software training set and testing set of calibration, and modify the YOLOv3 classifier, change the output of the network dimension, optimization of network parameters and accelerate the convergence of the model, by analysing the experimental result shows that using YOLOv3 network, can realize the four directions to set an underwater vehicle motion behavior recognition, and ensure the accuracy and speed, the foundation for subsequent underwater vehicle based on visual collaboration.

Keywords: YOLOv3, Neural Networks, Underwater Vehicle Detection, Target Recognition

1. Introduction

As an important tool for human beings to explore the ocean, the autonomous cooperation of underwater vehicle has attracted the attention of researchers from all over the world [1,2]. However, compared with mobile robots and unmanned aerial vehicles (UAVs), autonomous cooperation of underwater vehicle is faced with such problems as the lack of real-time underwater communication and the limited sensors for sensing other cooperative objects.

The underwater information transmission between the underwater vehicle and the mother ship (boat), other underwater unmanned vehicles, submariners or buoys is mainly through underwater acoustic communication, so as to realize the underwater wireless transmission of data, information and intelligence [3]. But belongs to the typical weak communication because of the underwater acoustic communication, the communication latency, belt width, and in real ocean environment affected by
water temperature, reverberation is usually poor communication reliability (intermittent), which severely restricts the underwater vehicle information interaction between frequency and the amount of data, this leads to underwater vehicle cannot real-time collaboration between each other.

2. Method of Target Detection and Recognition: YOLOv3 Network

To realize the recognition and perception of the cooperative object of the underwater vehicle by vision, the object detection of the underwater vehicle should be solved first. The traditional target detection algorithm extracts the geometric and statistical features of the target through artificial design, while deep learning uses the network to select the learning image features for expression. The target detection mainly includes two key tasks: target classification and target positioning. The target classification needs to give the category label to all the regions of interest in the input image. The target positioning needs to find out all the positions including the target in the input image, the surrounding box of the output target object or the position of the target at the pixel level, and divide it out.

Target detection field, the more the more tired of using convolution neural network, is now a mature network mainly include: Girshick [4] the R-CNN network, put forward by Selective Search algorithm to generate the candidate regions, by convolution method to extract features, using SVM to complete classification; The Fast-RCNN network model proposed by Girshick [5]. adopts softmax for classification and Smooth L1 Loss for regression box detection. Ren [6] proposed Faster-RCNN, and introduced RPN (Region Proposal Network) to directly predict the candidate box, so as to complete end-to-end detection and further improve the target detection accuracy. YOLO is an effective target detection model without region Suggestions, which abandons the original generation of candidate regions[7]. The network can simultaneously conduct classification and location prediction of multiple categories of targets, and can directly train the whole network end-to-end. YOLOv2, the change the VGG-16 of the network structure, and use a small amount of calculation of Darknet-19, moreover also through standardizing the Batch (Batch Normalization, BN), and training of multi-scale strategy improve detection accuracy[8]. YOLOv3 using logistic regression to predict, with reference to residual network residual Darknet-53 network structure of the network, improve the prediction precision, to strengthen the recognition of small objects. Meanwhile, SSD [9] combined the regression thought of YOLO and the anchor concept of Faster-RCNN, and adopted the multi-scale characteristic.

After deep convolutional neural network was applied to the behavior recognition task, Simonyan [10] first proposed the two-stream ConvNets model, which included Spatial Stream and Temporal Stream modeling. Based on the dual-flow structure, Wang [11] further subdivided the time-domain flow into local time-domain flow and global time-domain flow. At present, behavior recognition mainly focuses on human behavior, and researches on robots, especially underwater vehicle, are relatively rare.

In this paper, the mature YOLOv3 network structure is mainly used to study the behavior recognition and perception of cooperative objects of underwater vehicle.

2.1. YOLOv3 Network Structure

YOLOv3 adopts the method of feature fusion and multi-scale detection to effectively improve the accuracy and speed of target detection. The YOLOv3 network structure has been developed from darknet-19 of YOLOv2 to darknet-53. The pooling layer and full connection layer in YOLOv2 were removed, and the dimensional transformation of the tensor was realized by changing the step size of the convolution kernel in the process of forward propagation. YOLOv3 network borrows from residual network, and each residual component has two convolutional layers and a shortcut link, as shown in Figure 1(b). Adjusted the network structure is made of Darknet-53 \((2 + 1 + 1 + 2 \times 2 \times 2 + 1 + 1 + 8 + 8 \times 2 \times 2 + 1 + 1 + 4 = 53)\) in accordance with the order number, the last is Connected fully Connected layer and convolution, a total of 53), as shown in Figure 1(a), add residual module in network, is beneficial to solve the problem of deep network of gradient, the diagram 1,2,8,8,4 represent several repeated residual module, in the whole network structure, no pooling and full
connection layer. The down sampling of the network is realized by setting the convolution stride as 2. After passing through this convolutional Layer, the size of the image will be reduced to half. However, the realization of each convolutional Layer includes Con2d-Layer, BN-Layer and LeakyReLU-layer, as shown in Figure 1(c). Each residual layer contained Convolutional (1×1) and Convolutional (3×3), as shown in Figure 1(d).

![Network Structure Diagram](image)

**Figure 1.** YOLOv3 network structure. (a) Darknet-53 network structure ;(b)Structure diagram of residual network; (c)Implementation structure diagram of the convolution layer; (d) Structure diagram of the Residual layer

With the change of the number and scale of the output characteristic graphs, YOLOv3 continues YOLOv2 to obtain the size of the prior boxes by K-means clustering, and sets three kinds of prior boxes for each sampling scale, and altogether coclustering out the prior boxes of nine sizes. For a 416×416 input image, three prior boxes are set in each grid of the feature map of each scale, and there are a total of 13×13×3 + 26×26×3 + 52×52×3 = 10647 predictions. Each prediction is a (4+1+80)= 85-dimensional vector containing frame coordinates (4 values), frame confidence (1 value), and the probability of the object category.

### 2.2. Classifiers and Loss Functions

YOLOv3, aiming at the problem of high detection rate of small targets, adds multi-level prediction from top to bottom, and outputs three feature graphs of different scales.YOLOv3 sets 3 candidate boxes for each grid cell, each candidate box has 5 basic parameters (x, y, w, h, confidence), and has the probability of 80 categories, so the dimension of network output is 3×(5+80)=255. In this paper, the motion behavior target detection of underwater vehicle is tested in four categories: go_forward, go_retreat, turn_left and turn_right. The tensor of the output dimension is 3×(5+4)=27. Therefore, the classifier of the model is modified on the basis of the original YOLOv3, and the output of the network model is modified to a tensor of 27 dimensions.

The loss function of YOLOv3 was changed on the basis of YOLOv2. YOLOv3 eliminated softmax and replaced it with logistic, and replaced the classification loss with binary cross entropy. The smaller the cross entropy, the closer the two probability distributions are as follow.

\[
H(p, q) = - \sum p(x) \log q(x)
\]  

(1)

where p is the distribution of the correct answer, q is the distribution of the prediction, and log base e.

The loss function of YOLOv3:
\[
\text{Loss} = \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^b \left[ (x_i - \hat{x}_i)^2 + (y_j - \hat{y}_j)^2 \right] + \lambda_{\text{obj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^b \left[ \sqrt{(\hat{w}_i - \hat{w}_i)^2 + (\hat{h}_i - \hat{h}_i)^2} \right] - \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^b \left[ \log(C_i^j) + (1-C_i^j) \log(1-C_i^j) \right]
\]

\[
\sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^b \left[ \hat{C}_i^j \log(\hat{C}_i^j) + (1-\hat{C}_i^j) \log(1-\hat{C}_i^j) \right] - \lambda_{\text{obj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^b \left[ \hat{C}_i^j \log(C_i^j) + (1-\hat{C}_i^j) \log(1-C_i^j) \right]
\]

\[
\sum_{i=0}^{S^2} \sum_{j=0}^{B} \left[ \hat{P}_i^j \log(P_i^j) + (1-\hat{P}_i^j) \log(1-P_i^j) \right]
\]
In order to speed up the training of the network, batch=32 and subdivisions=8 were set by loading the darknet-53 pre-training model, so as to randomly extract 32 training samples from the training set in each iteration, divide them into 8 groups and send them to the network for training, so as to reduce the memory pressure. The learning rate is 0.001, because if the learning rate is too large, the weight update speed will be too fast, and the optimal value will be easily passed. On the contrary, if the learning rate is too small, the update rate will be slow and the efficiency will be low. On this basis, the training of the model was carried out by using the data set of underwater vehicle motion behavior recognition and the training log file (train_yolov3.log) was saved. The loss curve of YOLOv3 algorithm is shown in Figure 2. The y-coordinate avg_loss represents the change curve of the loss function of the network structure in the training process, and is expected to approach 0 at last. Since the value of loss function was large in the first 700 iterations, when the avg_loss curve was drawn, the change curve was drawn from 700 iterations, and the change curve of loss function was observed. It can be seen that after the completion of 700 iterations, Loos dropped below 0.6, and when the training iterations reached 1000, Loos dropped to 0.3. As the training continued to iterate, the Loos value continued to decline. After 8,000 iterations, the Loos value basically did not decrease and became stable. It can be seen that the YOLOv3 network has a rapid loss reduction and fast network convergence speed in the training of underwater vehicle motion behavior recognition data set.

The IOU variation curve of YOLOv3 algorithm is shown in Figure 3. The value of Avg IOU represents the ratio of intersection and union between the candidate box and the real marked border in the current training iteration. The value of Avg IOU should be closer to 1.0. According to the Avg IOU change curve, as the number of iterations increases, the value of Avg IOU keeps rising. When the number of iterations reaches 4000, the value of Avg IOU basically approaches to the expected value 1.0. The rectangular box and the target coincide well, and the behavior performance of the model detection underwater vehicle is better. After the training of underwater vehicle motion behavior recognition training set is completed by using YOLOv3 network, the test set is input into the trained network for test verification, and the detection result of YOLOv3 on the underwater vehicle behavior in the test set is obtained, as shown in Figure 4.

According to the observation and recognition results, YOLOv3 has a high recognition accuracy for the motion behavior of the 4 underwater vehicle designed, but for the image with fuzzy target pixels, the phenomenon of missing and miscalculation will occur. In the underwater environment, due to the influence of water quality and impurities on the surface of the water, the objects far away from the fixed camera will be missed and misdetected. The identification results of YOLOv3 in test sets were statistically analyzed, and the accuracy and recall rates of the detection results were statistically analyzed, as shown in Table 1.
Table 1. Test results

| Project      | Accuracy rate % | Recall rate % | Average accuracy rate % |
|--------------|-----------------|---------------|-------------------------|
| Go_forward   | 89.46           | 85.71         | 86.33                   |
| Go_retreat   | 73.638          | 71.39         | 71.33                   |
| Turn_left    | 82.19           | 80.00         | 81.78                   |
| Turn_right   | 76.87           | 72.73         | 73.56                   |

According to the results, the YOLOv3 network has a good effect on the detection and recognition of the four motion behaviors set by the underwater vehicle, with the average detection speed of the network reaching 0.039s.

4. Conclusion

This paper introduces in detail the detection and recognition process of four kinds of underwater vehicle motion behavior based on YOLOv3 network model, including modifying classifier, optimizing network parameters, and collecting and calibrating experimental data sets. The experimental results show that the YOLOv3 network has achieved good results in detecting and recognizing the four behaviors of underwater vehicle. In the next step, the model will be further improved, and the number and quality of sample training will be increased to optimize the model. It is expected to realize real-time recognition of the motion behavior of underwater vehicle in the underwater environment, improve the accuracy of detection, and lay a foundation for the subsequent cooperation of underwater vehicle based on vision.

References

[1] Feng X S, Li Y P. Thirty years evolution of SIA’s unmanned marine vehicles. Chinese Science Bulletin, 2013, 58 (S2):2-7.
[2] Feng X S, Li Y P, XU H L. The next generation of unmanned marine vehicles dedicated to the 50 anniversary of the human world record diving 10912m. Robot, 2011, 33(01):113-118.
[3] Chen Q. Unmanned underwater vehicle. Beijing: Nation Defense Industry Press, 2014.
[4] 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):142-158.
[5] Girshick R. Fast R-CNN. Proceedings of the IEEE international conference on computer vision, 2015:1440-1448.
[6] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Neural information processing systems, 2015:91-99.
[7] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:779-788.
[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition,2017:7263-7271.
[9] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. European conference on computer vision. Springer, Cham, 2016:21-37.
[10] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advance in neural information processing systems, 2014:568-576.
[11] Wang L, Ge L, Li R, et al. Three-stream CNNs for action recognition. Pattern Recognition Letters, 2017, 92:33-40.