International Core Journal of Engineering Volume 7 Issue 5, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

Water Area Object Detection based on YOLO-Fusion Network

Yuhui Yang[a], Bin Li

Logistics Engineering College, Shanghai Maritime University, Shanghai 201306, China.

[a]1119032324@qq.com

Abstract

In order to improve the efficiency of controlling tourists' private launching behavior in dangerous waters such as seashores and reservoirs, a small target detection network in waters optimized based on the You Only Look Once version 3 (YOLO v3) network structure is proposed. On the basis of YOLO v3, a side road network composed of feature fusion modules is designed. The side road network effectively improves the transmission and utilization of feature information between network layers, especially for fusion interaction between the network shallow feature information and deep information. At the same time, the Spatial Pyramid Pooling Module (SPP Module) is also added to perform multi-scale aggregation of the output features of the main network and the side road network to enhance the characterization ability of the feature map. In this paper, the optimized network is called YOLO-Fusion. After testing, the average detection accuracy of YOLO-Fusion is 89.55%, which is 2.45% higher than that of YOLO v3. It is based on test indicators such as precision and recall. There is also a significant improvement.

Keywords

Convolutional Neural Network; Object Detection; Feature Fusion; YOLO v3.

1. Introduction

The number of deaths caused by drowning accidents accounts for the top three deaths in accidents every year. According to statistics, the proportion of drowning accidents in the sea area accounts for 76%. Therefore, many coastal cities in my country have set prohibitions on beaches. And assign someone to manage it. However, due to the large area that needs to be controlled, the administrator cannot effectively control each area. When someone violates the prohibition or an accident occurs, they cannot stop or rescue them immediately. By means of computer vision, the designated area is detected and searched. When the target is detected, an alarm is issued to remind the management staff to pay attention, which can effectively improve the supervision and avoid the occurrence of accidents.

One of the earliest and most widely used object detection technologies in computer vision is pedestrian detection. Traditional target detection algorithms take Haar-like features (Haar) proposed by PAPAGEORGIOU et al. [1] and Histogram of Oriented Gradient (HOG) proposed by DALAL et al. [2] as typical representatives. The detection of the target by the class algorithm mainly has three steps: obtaining the candidate area through the sliding window, extracting the features of the candidate area, and then sending the extracted features to the classifier for classification.However, because the features of this type of feature detector are designed based on the prior knowledge and subjective consciousness of the designer, the overall generalization effect of the algorithm is not good, and there are certain restrictions on the detected scene. In addition, the candidate area is generated based on a sliding window. This strategy is not targeted and has high time complexity and computational redundancy.At present, the target detection algorithm is more based on deep learning. In the ILSVRC competition in 2012, Alex et al. [3] won the championship with the AlexNet model

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

and made deep learning as a synonym for neural networks well known. Then Ross et al. in 2014 [4] The proposed R-CNN (Region Convolutinal Neural Network) network model and the YOLO (You Only Look Once) network model proposed by Joseph et al. [5] in 2016 have made deep learning more concerned in the field of computer vision.The principle of deep learning is mainly to automatically learn the features of training samples through convolutional neural networks. Compared with manual design of features, the features obtained by independent learning of neural networks will be more suitable for the target, so the detection accuracy of detection algorithms based on deep learning Higher and overall robustness is more robust.The R-CNN series and the YOLO series respectively represent two deep learning algorithms with different detection methods. The R-CNN series is a target detection algorithm based on the recommendation of candidate regions. The entire detection process can be divided into two parts. It first selects foreground regions that may have targets on the feature map, and then performs feature extraction on these regions and based on the extracted features The output of the final test result.The YOLO series algorithm is based on an end-to-end linear regression target detection algorithm, which is different from the R-CNN series network YOLO directly performs feature extraction operations on the input pictures and makes predictions based on the extracted features, and completes feature extraction and Inspection work.In summary, the advantage of the R-CNN series algorithm is that the two-stage detection process makes its detection accuracy higher, and the YOLO series algorithm makes its detection speed faster due to the integrated detection network structure.

In this paper, based on the actual detection scenario, an algorithm is required to quickly complete the detection of the input image or video, so that when the above situation occurs, an alarm can be issued to remind the manager immediately.Therefore, this paper selects the YOLO v3 algorithm, which has a faster inference calculation speed, as the basic algorithm, and optimizes the network structure based on it to make it more suitable for the detection scenarios described in this paper.The optimization idea for the original network is mainly to strengthen the use of features between layers, so that the YOLO detection module can use more shallow representation information to search for targets, solve the problem of missing detection of the original network structure, and improve the overall detection accuracy.

2. YOLO v3 object detection algorithm

YOLO v3 uses the deep convolutional neural network Darknet53 as the backbone network, and is equipped with three YOLO detection modules. The overall network structure of YOLO v3 is shown in the following figure:

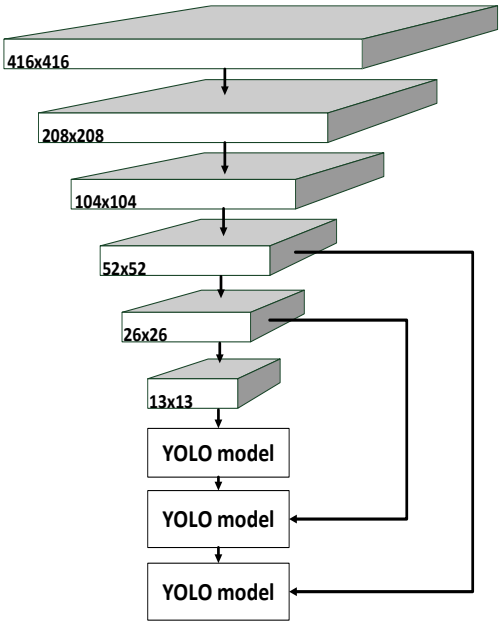


Fig. 1 YOLO v3 network structure

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

The Darknet53 feature extraction network as a whole contains five scale residual modules to extract feature maps of different dimensions. After the feature extraction operation, the extracted features will be input into the three YOLO modules to integrate the feature information again and the YOLO layer Detect the target.Each module in the feature extraction network is composed of multiple identical residual units. The structure of the residual unit is shown in the figure below:

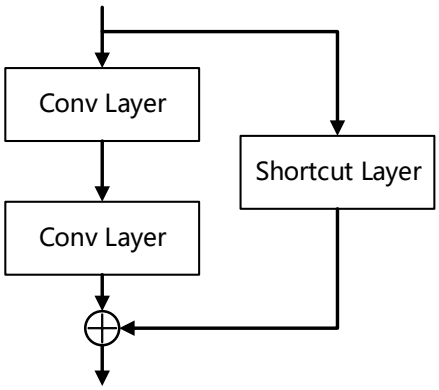


Fig. 2 Residual unit structure

Each residual unit is composed of two convolutional layers and a short-hop connection layer. The convolutional layer is responsible for extracting features, and the short-hop connection layer is responsible for identity mapping. The input of the residual unit is mapped to the output, avoiding the The problem that the gradient disappears during training when the network is too deep. After the feature extraction operation, at the end of the network are three YOLO detectors of different scales, which correspond to the detection tasks of large, medium and small targets respectively. The overall network design of YOLO v3 adds a feature fusion structure, which combines the characterization information extracted from the shallow network with the semantic information of the deep network, and the feature map obtained by the fusion is then detected by the YOLO detector.

The YOLO detector is different from RCNN for target classification and positioning based on the candidate regions on the feature map. It directly performs the detection task on the complete feature map. First, the YOLO detector divides a grid on the input feature map, that is, divides a detection task into different grid areas. When the center point of the target is in a certain grid, the grid is responsible for detecting the target. Each grid is equipped with three anchors. Based on the size of these three anchors, three Brounding boxes are generated to predict the category and position of the target. Each Brounding box combines the category confidence score and the position confidence score to form the final prediction confidence score. The Brounding box with the highest score is the final prediction result, and the remaining two Brounding boxes are deleted. The category confidence score is $Pr(class_i|object)$, that is, the probability that the current target is the i-th category; the position confidence score is composed of five parameters (x, y, w, h, c), and (x, y, w, h) is the Brounding box The coordinate information, c is the area intersection ratio between Brounding box and Ground truth, and its confidence score is IOU_{pred}^{truth}. In summary, the formula for calculating the overall confidence score of the Brounding box is:

$$C = Pr(class_i|object) * Pr(object) * IOU_{pred}^{truth} = Pr(class_i) * IOU_{pred}^{truth} \quad (1)$$

In the above formula, Pr(object) is the predicted probability of foreground and background. When there is a target in the Brounding box, its value is 1, otherwise it is 0. Finally, the loss function formula is used to calculate the error value between the confidence score of the Brounding box and the marking information of the Ground Truth, and the network parameters are updated according to the back propagation calculation of the error value. The YOLO v3 detector uses the Logistic classifier to calculate the center point coordinates, length and width, foreground and background discrimination and target category prediction errors of the Brouding box.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

3. YOLO network based on strong feature fusion

Combined with the actual application scenarios of this article to detect pedestrians on the sea or on the beach, there are two areas for improvement in the YOLO v3 network. One is that the overall structure is a deep convolutional neural network, and the entire network has 5 downsampling operations. After the down-sampling operation, the effective pixel area of the target will gradually decrease, and the human target in the sea is a small target, which will lose more characteristic information, which is not conducive to the detection of the detector. Second, the distribution of detectors is unreasonable in this application scenario. The original network structure has three types of detectors: large, medium, and small, but in the actual monitoring overhead screen, the targets (tourists) in the sea are all small targets. There is no situation where a large area occupies the frame, so the large-scale detector is not suitable for the scene described in this article [6-8].

In view of the above problems, this paper designs a strong feature fusion network based on the original network structure. This network enhances the feature transferability between the residual module and the residual module, and also enhances the fusion of shallow features and deep features. The structure of the strong feature fusion network is shown in the figure below:

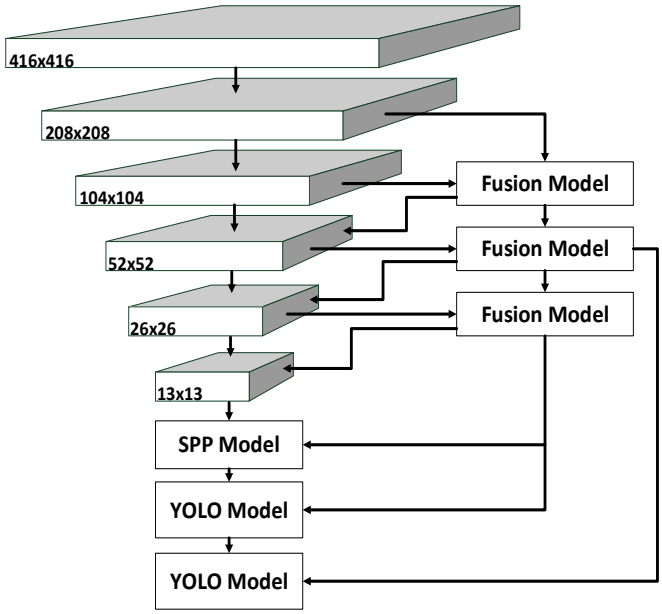


Fig. 3 Strong feature fusion network

First, three fusion modules (Fusion Model) are added to the structure to construct a side road feature fusion network. The main function of the side road network is to collect and transmit the features extracted by the residual module of each scale, and improve the utilization of feature information. The fusion module is composed of a concat layer and a maximum pooling layer. The concat layer is responsible for channel fusion of the feature maps extracted by the two residual modules. The fused feature maps are passed to the next scale through the downsampling operation of the maximum pooling layer Residual module and fusion module. As shown in Figure 3, the fusion module corresponding to the 104 scale fuses the output feature maps of the last residual module in the 208 scale and the 104 scale. The fused feature map is input into the first residual module of 52 scale to provide it with the comprehensive features of the upper network. At the same time, it will also be passed to the output of the next fusion module and the last residual module of 52 scale for feature fusion to continue to enrich the feature information of the side road network. At the same time, compared to the original network that directly transfers the shallow feature information from the backbone network to the YOLO detection module, in this article, it is transmitted through the side road network. The 52-scale and 26-scale residual modules of the original network are in the lower

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

layer of the network. The input image data has undergone a certain convolution operation, and its feature information is not representative, but more semantic features. The feature information in the shallow part of the network more reflects the morphological characteristics of the target, such as texture, size, color, and location. At the same time, for the small target in this article, as the size of the feature map is reduced, the area of the target area is also reduced, and the feature information that the network can extract gradually decreases. Therefore, the feature fusion method adopted by the original network can transmit less effective information. In the optimized structure, the second and third fusion modules correspond to the feature maps of 52 and 26 scales respectively. In addition to the feature information of the residual modules of the corresponding scales, they also have the characteristics of shallow networks at the 208 and 104 scales. information. The features of these two fusion modules are respectively transferred to the 52 and 26 scale YOLO detection modules. Compared with before optimization, this method makes the feature information obtained by the detector more comprehensive, and can combine shallow representation information and deep semantic information to detect targets [9-10].

Secondly, an SPP module is added at the end of the feature extraction network. The SPP module is composed of a maximum pooling layer, a route layer, and a concat layer. The input data of the SPP module is the feature map after the fusion of the Darknet network and the side road network. The feature map is scaled to different scales by using different size pooling cores. Then use the concat layer to fuse all the zoomed feature maps and output them to the subsequent YOLO detection module. The SPP module realizes the fusion of local features and global features. The feature map output by the SPP module aggregates multi-scale feature information, that is, integrates the output features of the feature extraction network, and strengthens the characterization ability of the feature map [11-13].Finally, according to the characteristics that the detection scenes are all small and medium-sized targets, the distribution of the original network detector is adjusted, the 13-scale detector is removed, and only the 26 and 52 small and medium-scale YOLO detectors are retained. In this paper, te YOLO algorithm using a strong feature fusion network is called YOLO-Fusion.

Experimental results and analysis

This experiment uses 2000 overhead pictures of tourists swimming in the sea or swimming pool crawled from the Internet, which are divided into training set and test set according to the ratio of 9:1. Since the size of the target detected in this article is small and the area of common targets is different, it is necessary to modify the hyperparameter anchors of the YOLO detector before YOLO-Fusion is trained to make the generated prediction box more suitable for the target. At the same time increase the value of iou. Using the k-means++ algorithm, multiple iterations to calculate the clustering values of the length and width of the target area, the final results are shown in the following table:

Table 1. Anchors clustering results

52x52			26x26		
(9,11)	(15,12)	(20,18)	(22,23)	(27,25)	(30,31)

YOLO-Fusion and YOLO v3 were trained in the same training environment, and the two models obtained were compared and analyzed using the test set. This paper uses confusion matrix as the test basis for model performance. The indicators obtained by confusion matrix test mainly fall into these three categories: TP (number of correct detections), FP (number of wrong detections), and FN (number of missed detections). According to the index data obtained by the confusion matrix test, the accuracy, recall, and comprehensive score (F1-socre) of the model can be calculated, which more intuitively reflects the detection performance of the model. The calculation formulas of the three indicators are:

$$precision = \frac{TP}{TP+FP} \quad (2)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

$$recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1 - score = 2 \times \frac{precision \cdot recall}{precision + recall} \tag{4}$$

At the same time, set the accuracy index as the Y axis and the recall index as the X axis. The enclosing area of the curve drawn according to the corresponding relationship between the two index values can represent the average precision (AP) of each type of target and the mean average precision (mAP) of the model detection performance. Since the target detected in this paper is a single class, the AP value is the same as the mAP value. The comparison of the test results of the two models on the test set is shown in the following table:

Table 2. Comparison results of model checking performance

	precison	recall	F1-score	IoU	AP
YOLO-Fusion	88.46%	84.03%	86%	75.64%	87.15%
YOLO V3	84.64%	81.36%	83%	71.05%	82.98%

From the comparison data in the above table, it can be seen that the detection performance of YOLO-Fusion is better than that of YOLO v3, especially from the two data of recall and IoU. Recall reflects the full retrieval performance of the model for the target. The strong feature fusion network enhances the use of shallow representation information and improves the efficiency of feature transfer. Compared with YOLO v3, YOLO-Fusion has increased the recall rate by 2.67%, and to a certain extent solved the problem of the original network missed detection targets. In addition, the IoU index of YOLO-Fusion is also significantly improved compared to YOLO v3, with an increase of 4.59%. The IoU index reflects the accuracy of the model for the location of the target area. In addition to relying on the shallow information provided by the optimized side road network for better positioning, the added SPP module aggregates global and local feature information by scaling the feature map, So that the model can accurately fit the area where the target is located from the macro and micro. At the same time, after clustering calculation, the size of the anchors of the detector is consistent with the actual size of the target, and the size of the predicted area generated on the basis of the anchors is consistent with the area of the actual target area.

Draw the PR curves of the two models according to the data obtained from the test, and compare the performance of the two models more intuitively:

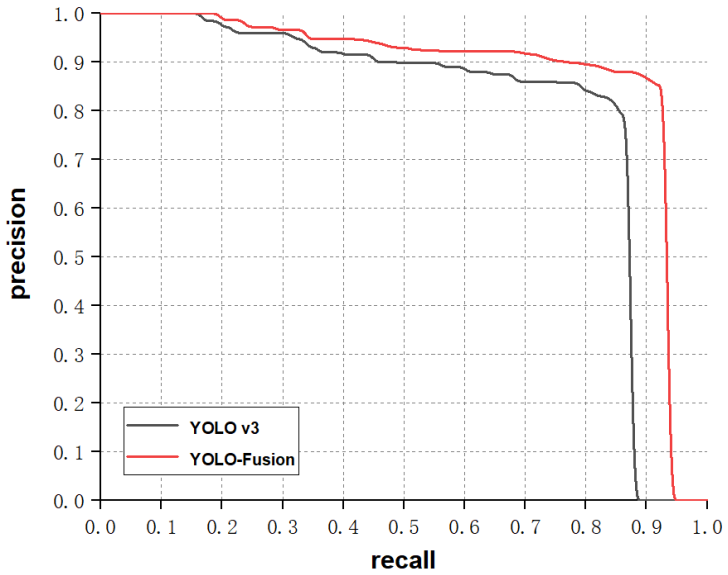


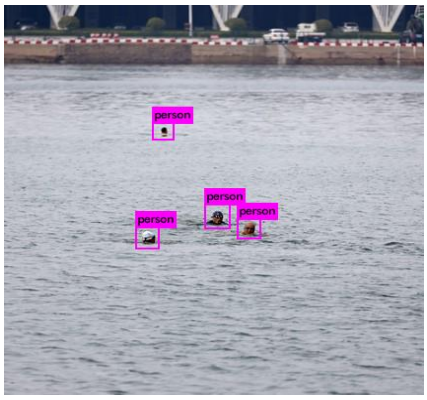
Fig. 4 Comparison of PR curves

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

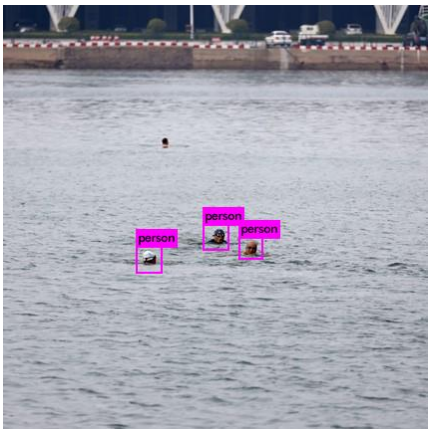
The actual detection effect comparison between YOLO-Fusion and YOLO v3 in the test set is shown below:

(a) YOLO-Fusion

(b) YOLO v3



(c) YOLO-Fusion



(d) YOLO v3

Fig. 5 Actual detection effect

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0013

From the control group a and b, it can be clearly seen that YOLO v3 is not as good as YOLO-Fusion in detecting dense targets in the distance, and there are some missed detections. YOLO-Fusion's detection effect is better in the overall detection situation. At the same time, it can be more intuitively observed from the c and d control groups that YOLO-Fusion is better than YOLO v3 for detecting small distant targets.

4. Conclusion

This paper designs a YOLO-Fusion network structure based on strong feature fusion, and applies it to the detection of tourist targets in the dangerous waters control area. The feature fusion module is constructed by using the concat layer and the maxpooling layer. Three feature fusion modules are used in the entire network structure, and these three feature fusion modules are constructed as a side road network. YOLO-Fusion relies on the side road feature fusion network to strengthen the application of the deep network to the shallow network features. At the same time, the SPP network module was added before the YOLO detection module, and the output characteristics of the main network and the side road network were aggregated at multiple scales. The local feature information and the global feature information are re-aggregated together to make the feature map features input to the YOLO detection module more comprehensive. The final experimental results show that compared with YOLO v3, YOLO-Fusion has significantly improved test indicators such as recall rate and detection accuracy, alleviating the problem of YOLO v3's missed detection of small targets. The next step of this paper will increase the detection speed while ensuring the detection accuracy, and design a complete detection system to improve its practicability.

References

[1] Papageorgiou, Oren and Poggio, A general framework for object detection[J], International Conference on Computer Vision, 1998.

[2] N Dalal, B Triggs. Histograms of Oriented Gradients for Human Detection [J]. IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2005, 1(12): 886-893.

[3] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C], International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[4] Girshick, Ross, et al. “Rich feature hierarchies for accurate object detection and semantic segmentation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi You Only Look Once:

[6] Unifified, Real-Time Object Detection Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[7] Huicheng Zheng,Jiajie Chen,Lvran Chen,Ye Li,Zhiwei Yan. Feature Enhancement for Multi-scale Object Detection[J]. Neural Processing Letters,2020,51(1).

[8] Wang Haiyun, Wang Jianping, Luo Fuhua. Research on Surface Defect Detection of Metal Plate and Strip Based on Faster R-CNN with Multi-level Features[J/OL]. Mechanical Science and Technology: 1-9[2020-04-28].https://doi .org/10.13433/j.cnki.1003-8728.20200024.

[9] Chen Han, Zhou Qiang. A method for judging people falling into the water based on reflection image detection[J]. Computer Knowledge and Technology,2018,14(26):175-178+180.

[10] Hao Zhu,Wenping Ma,Lingling Li,Licheng Jiao,Shuyuan Yang,Biao Hou. A Dual–Branch Attention fusion deep network for multiresolution remote–Sensing image classification[J]. Elsevier B.V.,2020,58.

[11] Zhao Jumin,Zhang Chen,Li Dengao,Niu Jing. Combining multi-scale feature fusion with multi-attribute grading, a CNN model for benign and malignant classification of pulmonary nodules.[J]. Pubmed,2020.

[12] Qi Lin,Zhang Haoran,Cao Xuehao,Lyu Xuyang,Xu Lisheng,Yang Benqiang,Ou Yangming. Multi-Scale Feature Fusion Convolutional Neural Network for Concurrent Segmentation of Left Ventricle and Myocardium in Cardiac MR Images[J]. American Scientific Publishers,2020,10(5).

[13] Minwei Deng. Robust human gesture recognition by leveraging multi-scale feature fusion[J]. Elsevier B.V., 2020,83.