ISSN: 2688-8645

Multiple Object Tracking Algorithm based on Joint Target Detection

Lina Xia

School of Merchant Shipping, Shanghai Maritime University, Shanghai 201306, China

Abstract

Due to the limitation of manpower, it is difficult to realize real-time feedback of channel conditions by manual to avoid risks. Real-time video supervision of shipping vessels can achieve dynamic real-time tracking and positioning, especially for specific ships and waters, such as dangerous goods ships, Tracking and monitoring of limited areas has become an urgent problem to be solved. This paper proposes a multiple object tracking algorithm based on target detection, Using CenterNet and YOLOv3 as the basic detectors, combined with DeepSort, to achieve real-time detection and tracking of ships. Through a large number of video tests to compare the actual effects of CenterNet – DeepSORT and YOLOv3-DeepSORT, the experimental results show that the detection and tracking algorithm of CenterNet-DeepSORT as the main structure can solve the current needs of actual water traffic monitoring scenarios, and the actual accuracy and robustness are both in multiple scenarios. It is stronger than YOLOv3-DeepSORT and has strong real-time and scalability. The analysis and extraction of trajectory features provide an expanded basis for subsequent intelligent traffic supervision and other auxiliary functions, such as abnormal ship behavior detection.

Keywords

CenterNet; DeepSort; Target detection; Multiple object tracking.

1. Introduction

The research on ship detection first began in the United States' autonomously controlled unmanned ship project, which is equipped with a ship detection system that can detect and track ship targets at sea [1]. Yang Guang et al. [2] proposed region segmentation, extracting the features of different segmented regions and processing them with machine learning algorithms to detect ships in each region.Xiong Yongping. [3] improved NMS (Non Maximum Suppression) based on the YOLOv2 network to solve the problem of ship detection under complicated sea conditions and different weather conditions. Yue Bangzheng et al. [4] proposed a ship target detection method based on improved Faster R-CNN, which solved the problem of missed detection of small ship targets. Liu Bo et al. [5] used YOLOv3 for training, which can identify important parts of the ship while detecting the type of ship, greatly improving the detection accuracy of small targets and overlapping occluded targets.

Previous academic research on tracking has shown that visual target tracking methods can be divided into generative methods, Discriminative Method. The water traffic environment is more complicated, such as partially or completely occluded by adjacent ships. Due to factors such as posture and illumination changes, sudden scale changes and motion blur, it is very challenging to perform long-term and robust ship tracking [6-7]. Zhou Yong[8] applied the Kalman filter algorithm to dynamically track the trajectory of ships in the crossing bridge area, thereby realizing intelligent monitoring and statistics of inland river ships. Chen Zechuan et al.[9] used the difference of gray peaks to detect the position of the ship, and tracked the ship

ISSN: 2688-8645

based on the mean shift.H. Grabner et al.[10] proposed a new semi-supervised robust tracking method (Semi B) to reduce the tracking drift problem.

With the rise of deep learning, many researchers have introduced target tracking methods based on deep learning [11-12]. Hyeonseob Nam et al. [13] proposed a CNN-based multi-domain learning model MDNET, which separates multiple target independent information from the target and applies it to the task of target tracking. Martin Danelljan et al. [14] proposed a tracking method ECO (efficient convolution operators)that combines deep learning and correlation filters, to factorize convolution operations, and adopt a new model update strategy to avoid the problem of model drift, the average tracking accuracy is higher.

In summary, with the in-depth research and development of the theory in the field of deep learning, single target tracking algorithms have made great progress. Because of the complexity and diversity of visual scenes and the interaction between targets, the related application research of deep learning in the multiple object tracking problem is still insufficient, and target tracking based on generative network models and unsupervised learning is becoming more and more popular in academic and industrial fields. s concern.

This paper proposes a detection joint online multiple object tracking algorithm based on deep learning. Compared with the traditional YOLOV3-DeepSort algorithm used in target detection, the CenterNet algorithm performs target detection with less light interference and more accurate detection of dense ships. Simulation and experimental results show that the optimized CenterNet-DeepSoet algorithm improves the accuracy of target detection, while effectively reducing the rate of false detections and missed detections, tracking and recognition are accurate, more stable, and strong anti-shake, which further improves ship target detection s efficiency.

2. Target Detection

Through the rapid development of computer vision technology and artificial intelligence technology, among many target detection algorithms, deep learning-based methods have been widely used because they do not require feature engineering, are highly adaptable, and are easy to convert. Therefore, using the method of deep learning to detect the target of the predicted image is the current research hot spot and the direction of future development.

Target detection algorithms based on deep learning are generally divided into two categories: two-stage detection models and one-stage detection models [15].

The two-stage model has two stages for image processing. It is a region-based method with high accuracy but poor real-time performance. Mainly include R-CNN, Fast R-CNN, Faster R-CNN, Mask-RCNN and so on. The single-stage model does not have an intermediate region detection process, and obtains the result directly from the picture. It is a region-free method with higher real-time performance but lower accuracy compared to two-stage detection. Mainly include YOLO, YOLOv2, YOLOv3, SSD and so on.

With the emergence of Corner Net, target detection has entered the era of anchorless. The early anchorless detection models mainly include: Dense Box, YOLO, and the new detection models mainly include: Extreme Net, Center Net, FSAF, FCOS, etc.

2.1. Principle of Center Net Model

Center Net is an anchor-free detection network [16], and it does not require NMS as post-processing like YOLOv3. It improves the detection speed while achieving accuracy. Higher effect. The overall framework and loss function borrowed from Corner Net [17] to a certain extent, and made effective improvements.

The basic idea is to treat the object as a point represented by its center rather than a bounding box when building a model, find the center point through key point estimation, and return to

ISSN: 2688-8645

other attributes, such as the width and height of the bounding box, to generate a prediction box. Realized by quoting the Heat map, and calculate the true predicted value according to the Gaussian distribution area of the predicted point. The offset prediction is used to indicate the coordinate error caused by the rounding operation when the annotation information is mapped from the input image to the output feature map.

2.1.1 Predict the Target Center Point

The Center Net network framework is shown in Figure 1. The detection network is divided into two stages: the down-sampling stage and the center point generation stage. In the down-sampling stage, the full convolutional network DLA-34 is executed to sample the image and obtain the sampling feature map. Based on the down-sampled feature map, the center point of the target is then generated. In the center point generation stage, three branch networks are used to predict the bounding box of the target at the same time, including the heat map generation branch, the center offset regression branch and the Bounding Box Size Regression Branch.

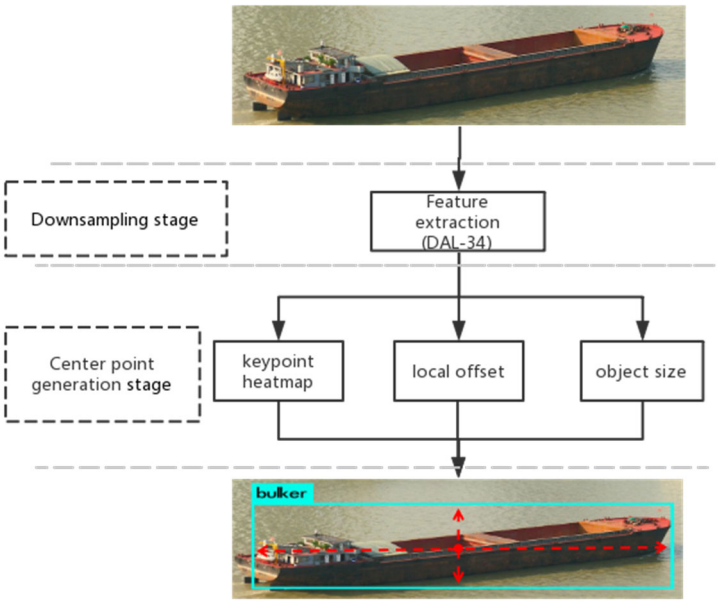


Figure 1. Center Net network framework diagram

The heat map generation branch is used to generate the center point of the target, and the number of channels of the heat map is the number of target categories that need to be detected. The center offset regression branch is used to correct the deviation between the true target center point of the branch generated from the heat map and the predicted target center point. The target bounding box size regression branch is used to predict the width and height of the target to generate a prediction box.

2.1.2 Heat Map Generation

In the heat map generation branch, first input an image with a size of 512×512, and use two-layer convolution to obtain a heat map containing key point information, and then use a 3×3 maximum pooling operation to obtain the possible position of the target center, avoiding doing NMS operation, simple and efficient filtering of repeated boxes. Secondly, k candidate points with the highest score are selected (the default value of k is 100), thereby determining the center point positions of 100 prediction boxes with the highest confidence.

ISSN: 2688-8645

As the target center point extracted by the network, the final heat map containing the target is output after being filtered by the threshold. The heat map can be expressed as Equation 1:

$$\hat{Y} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C} \tag{1}$$

Where W and H are the width and height of the input image respectively,R are the downsampling factors, and C are the detected key point categories (number of output feature channels, number of categories).

In the heat map, if the predicted value is $\hat{Y}_{x,y} = n$, it means the target center point with coordinates (x, y), and its corresponding classification number is n, if the predicted value is $\hat{Y}_{x,y} = 0$, it means no target point has been detected.

In the training phase, Gaussian Kernel is used to map key points to the feature map. The formula is as follows:

$$Y_{xyc} = exp\left(- \frac{(x - \tilde{p}_x)^2 + (y - \tilde{p}_y)^2}{2\sigma_p^2} \right) \tag{2}$$

In the above formula, $\tilde{p}_x$ and $\tilde{p}_y$ are the x, y coordinates of the center point of the real target, And σ is the standard deviation from the target size. The loss function of the heat map can be expressed as:

$$L_K = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - \hat{Y}_{xyc})^{\alpha} \, log(\hat{Y}_{xyc}) \\ (1 - Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} \, log(1 - \hat{Y}_{xyc}) \end{cases} \tag{3}$$

Where α and β are the hyperparameters of focal loss, and N is the number of key points.

2.1.3 offset regression

In the heat map generation branch, the sampling rate of the feature map is reduced by 4 times; therefore, when the feature map is remapped to the original image and an additional local offset is used for each center point, it will cause accuracy errors. In the offset regression branch, two layers of convolution are applied to obtain the offset feature map, which can be expressed as

$$\hat{O} \in R^{\frac{W}{R} \times \frac{H}{R} \times 2} \tag{4}$$

These two feature maps estimate the deviation of the x coordinate and y coordinate respectively. In the migration regression branch, the loss function can be expressed as:

$$L_{off} = \frac{1}{N} \sum_{p} \left| \hat{O}_{\tilde{p}} - \left(\frac{p}{R} - \tilde{p} \right) \right| \tag{5}$$

Among them, $\hat{O}_{\tilde{p}}$ is the predicted offset, and $\left(\frac{p}{R} - \tilde{p} \right)$ represents the offset between the center point of the original image and the center point of the down-sampled feature map.

2.1.4 bounding box size regression

ISSN: 2688-8645

In the bounding box size regression branch, two-layer convolution is applied to obtain the box size feature map. The feature map can be expressed as

$$\hat{S} \in R^{\frac{W}{R} \times \frac{H}{R} \times 2} \tag{6}$$

These two feature maps respectively predict the width and height of the relevant target. The loss function can be expressed as:

$$L_{size} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{S}_{pk} - S_k \right| \tag{7}$$

Where $\hat{S}_{pk}$ represents the size of the predicted frame, and S_k represents the size of the real frame.Use γ_{off} and γ_{size} to represent the hyperparameters of the loss function. The loss function in the training phase can be expressed as:

$$L_{size} = L_k + \gamma_{off} L_{off} + \gamma_{size} L_{size} \tag{8}$$

In summary, CenterNet is a minimalist representative of anchor-free detection methods. It is an end-to-end unique, simple and fast target detector, and it is more accurate than some anchor frame-based detectors. CenterNet provides three backbone network structures, as shown in Table 1:

Table 1. Three Network Structures of CenterNet

	cocoAP	FPS
Resnet-18	28.1%	142
DLA-34	37.4%	52
Hourglass-104	45.1%	1.4

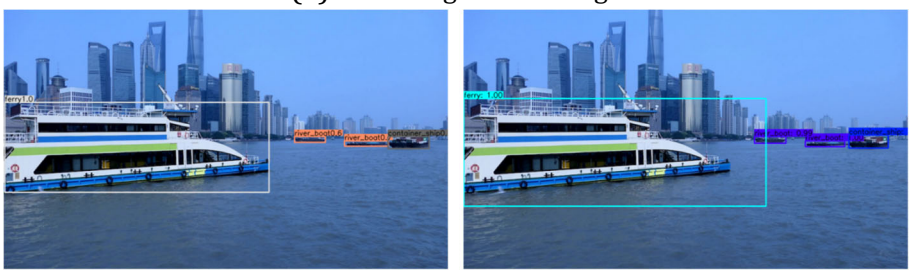
2.2. Target Detection Results

1)CenterNet has excellent detection effects for small targets at long distances and large targets at short distances; accurate detection under heavy load conditions; under different viewing angles: accurate positioning of the detection frame at flat and top angles; when the lighting environment changes and the target background is more complex, It is less affected by interference and has better robustness; effective capture of local features leads to larger detection frames; local multi-frame detection may occur when multi-purpose ships are detected; missed detection situations may occur when ships are too close, blocked, or dense Target group, etc.

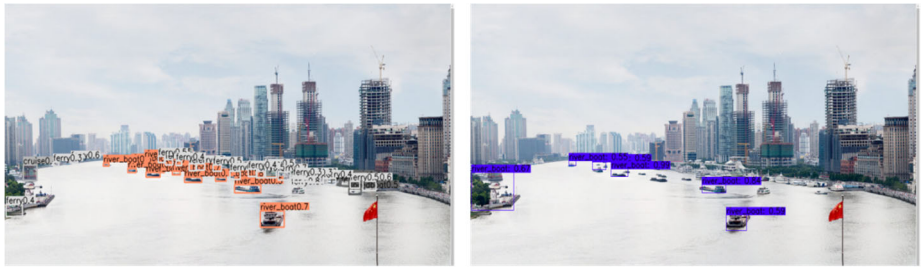
2)YOLOv3 for small and large targets as compared to the target detection CenterNet performance is poor, and there are many cases where missed detection frame inaccurate; flat, more accurate angle detection plan; poor overload is detected; for light There are many missed inspections when the environment changes, the background is more complicated, and the visibility is limited; missed inspections and false inspections occur when ships overlap. See Table 2 for details.

ISSN: 2688-8645

(a) Small target close range



(b) Flat angle of view



(c) Complex background

(d) Ship intensive

(e) Night

(f) Visibility changes

Figure 2. Comparison of target detection results

ISSN: 2688-8645

Table 2. Comparison of test results

	CenterNet	YOLOv3
Small target detection	better	Missed more
Large target detection	better	Redundant frame, inaccurate detection frame
Flat, top view angle	better	better
Lighting changes	better	Partially missed
Complex background	Great	Missed more
night	general	Missed more
Partially blocked, obstructed vision	There are missed inspections and redundant frames	Missed more

3. Multiple Object Tracking

Multiple Object Tracking, also known as MOT, is to track and extract multiple trajectories of interest in a video sequence, and assign tracking identification codes (Identity, ID) through time domain correlation to calculate their motion trajectory information. multiple object tracking is generally applied to longer videos, tracking multiple objects, involving multiple scenarios: appearance, departure, occlusion, etc. Unlike single-object tracking, multiple objects pay more attention to matching based on detected objects.

The general process of the MOT algorithm is: (1) The original frame of a given video; (2) Run the object detector to obtain the bounding box of the object; (3) For each detected object, calculate different features, usually Visual and motion characteristics; (4) similarity calculation step: calculate the probability that two objects belong to the same object; (5) correlation step: assign a digital ID to each object. As shown in Figure 3

Figure 3. MOT algorithm flow chart

Multiple object tracking is used in many fields, such as: automatic driving, behavior analysis, video analysis, intelligent traffic management, etc., in pursuit of reducing the negative effects of light changes, obstructions caused by obstructions, complex and similar object backgrounds, and smooth output.

In the current mainstream situation, multiple object tracking algorithms [18] are divided into two categories:

(1) multiple object tracking algorithm (CF, Correlation Filters) based on correlation filtering. The representative algorithm is a multiple object tracking algorithm based on kernel correlation (KCF)[19], which uses multiple threads to complete multiple single-object tracking. The algorithm samples the first frame of image, forms a cyclic matrix from the object area, and converts the calculation to the frequency domain through fast Fourier transform for solving, reducing the time complexity. By introducing ridge regression and nuclear techniques, the tracking speed and accuracy are effectively improved.

(2) multiple object tracking algorithm (TBD, Tracking-by-Detecton) based on object detection is currently a hot research topic in the industry, among which SORT and DeepSORT are well-known. Usually divided into multiple segments, the detector at the front end is responsible for

ISSN: 2688-8645

classifying the foreground target in each video frame, separating the background, and tracking at the back end, calculating the feature distance between targets in adjacent frames through Hungary, KM matching, etc., supplemented by IoU Wait for judgment, and perform target association and matching, so as to establish contact in the preceding and following frames.

3.1. SORT Tracking Algorithm

SORT [20] (Simple Online and Real-time Tracking)is the prototype of the currently popular TBD strategy tracking algorithm. The core is two algorithms: Kalman filtering and Hungarian matching. There are four basic components: target detector, state prediction, data association and track management update.

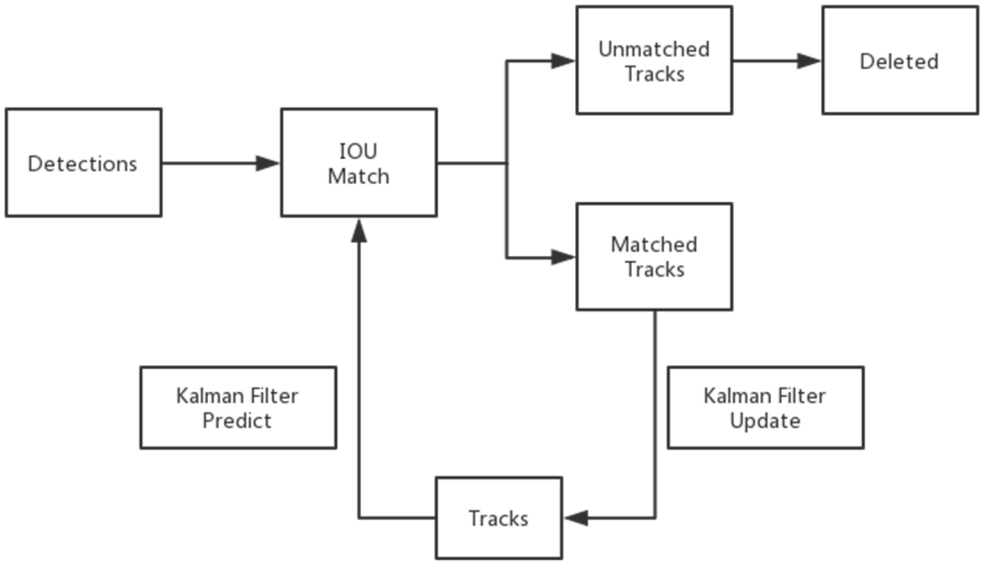


Figure 4. Description of SORT algorithm flow

The SORT algorithm as a whole can be divided into two parts, namely the matching process and the Kalman prediction and update process. The Kalman filter is used to predict, and then the Hungarian algorithm is used to match the predicted tracks with the detections in the current frame. Finally, Karl Mann filter updates the trajectory.

3.2. DeepSORT Tracking Algorithm

DeepSort (Simple Online And Real-time Tracking With A Deep Association Metric) is an improved algorithm of Sort [21].The optimization is mainly based on the cost matrix in the Hungarian algorithm. It did an additional cascade match before IOU Match, using appearance characteristics and Mahalanobis distance. The SORT algorithm uses Kalman filter to calculate data relevance frame by frame and uses the Hungarian algorithm to measure the relevance. This simple algorithm achieves good performance at high frame rates. However, because the surface features of the detected target are ignored, SORT can only work effectively when the uncertainty of state prediction is low. In actual situations, there are frequent tracking identification code loss. In DeepSORT, a convolutional network is used to extract the corresponding The feature map in the detection frame is judged whether it is the same target by calculating the distance between the feature tensors in the previous and subsequent video frames.In addition, in order to combat target loss caused by long-term occlusion, appearance information is added.

DeepSORT uses Mahalanobis distance and Cosine distance to calculate motion information and appearance information, respectively, as the similarity metric between motion features and depth features in the detection frame, and the overall similarity is obtained by weighted average It uses recursive Kalman filtering and cascade matching (Matching Cascade) to first

ISSN: 2688-8645

match the target of the track with higher recent activity frame by frame to improve the robustness of target tracking.

The Kalman filter is used for trajectory prediction, the Hungarian algorithm is used to match the predicted tracks with the detecions in the current frame (cascade matching and IOU matching), and finally the Kalman filter updates the trajectory.

The algorithm flow of DeepSort is shown in Figure 5:

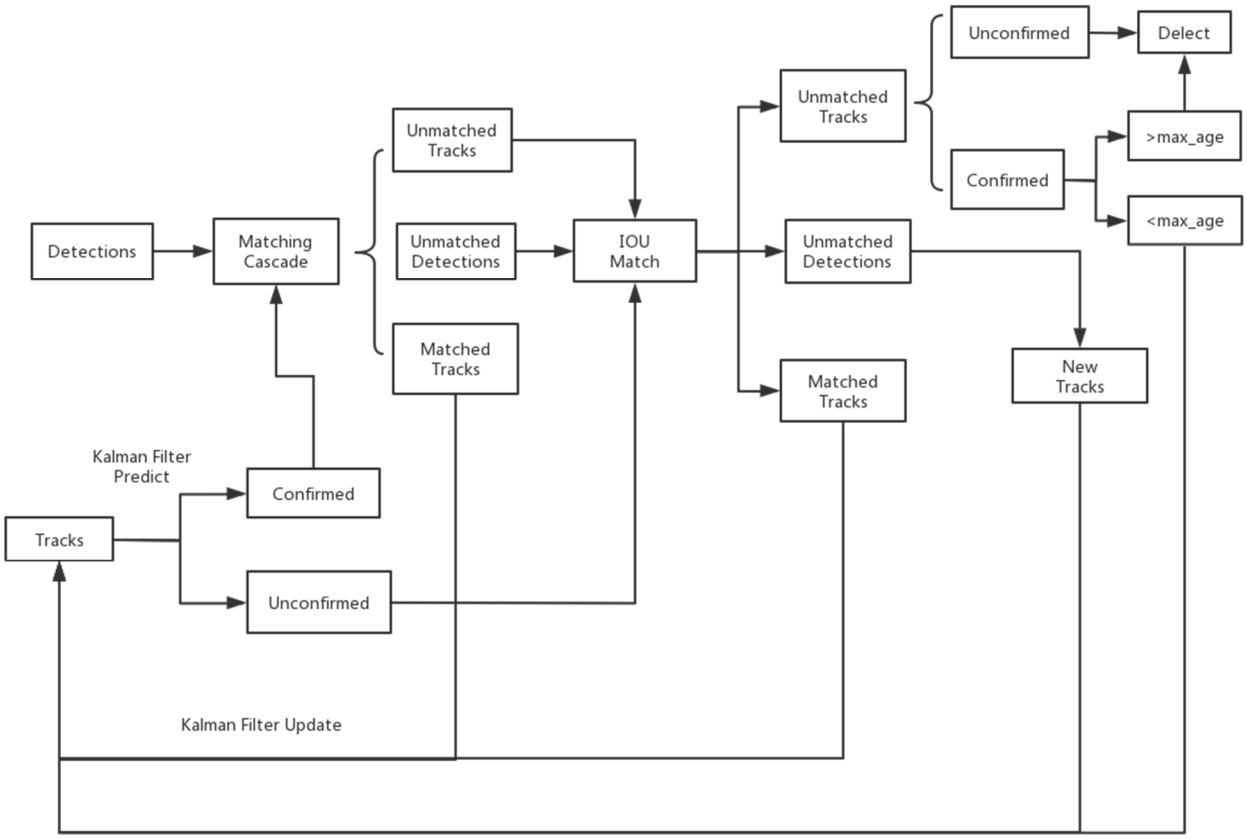


Figure 5. DeepSort SORT algorithm flow chart

Compared with the Sort algorithm, the integration of appearance information reduces the number of identity conversions by about 45 %. By using the deep convolutional network to extract target features as the matching standard, the robustness against occlusion and loss is increased, and the track life is greatly improved.

3.3. Real-time Tracking of Multiple Targets

As already on target detection and tracking algorithm we made related presentations, design and implement a multiple object tracking algorithm for ship surveillance, mainly CenterNet as a backbone network CenterNet-DeepSORT, while adding to YOLOv3 as a backbone network YOLOv3-DeepSORT were compared. The flowchart is shown in Figure 6:

3.4. Test Results

The research in this paper is oriented to ship monitoring. Therefore, a tracking test is carried out on the ship navigation video for CenterNet-DeepSORT and YOLOv3-DeepSORT in order to analyze and compare the performance of the two algorithms.

Through joint debugging to test the video tracking effect, part of the output is shown in Figure 7.

ISSN: 2688-8645

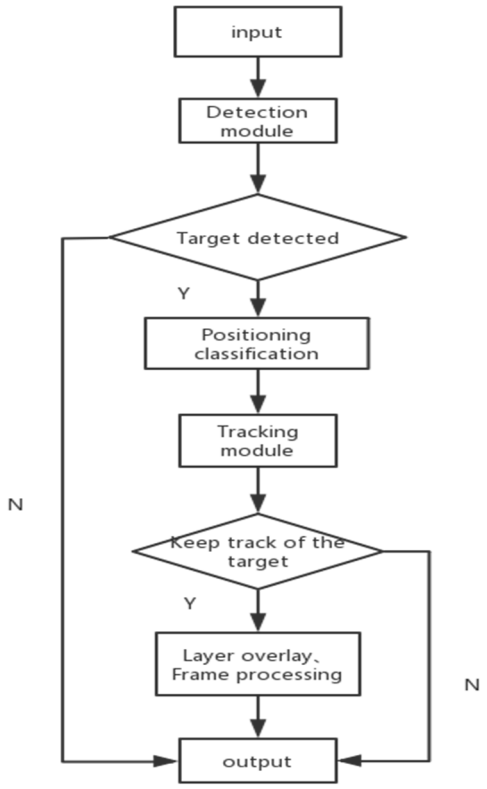


Figure 6. Flow chart of multiple object tracking algorithm

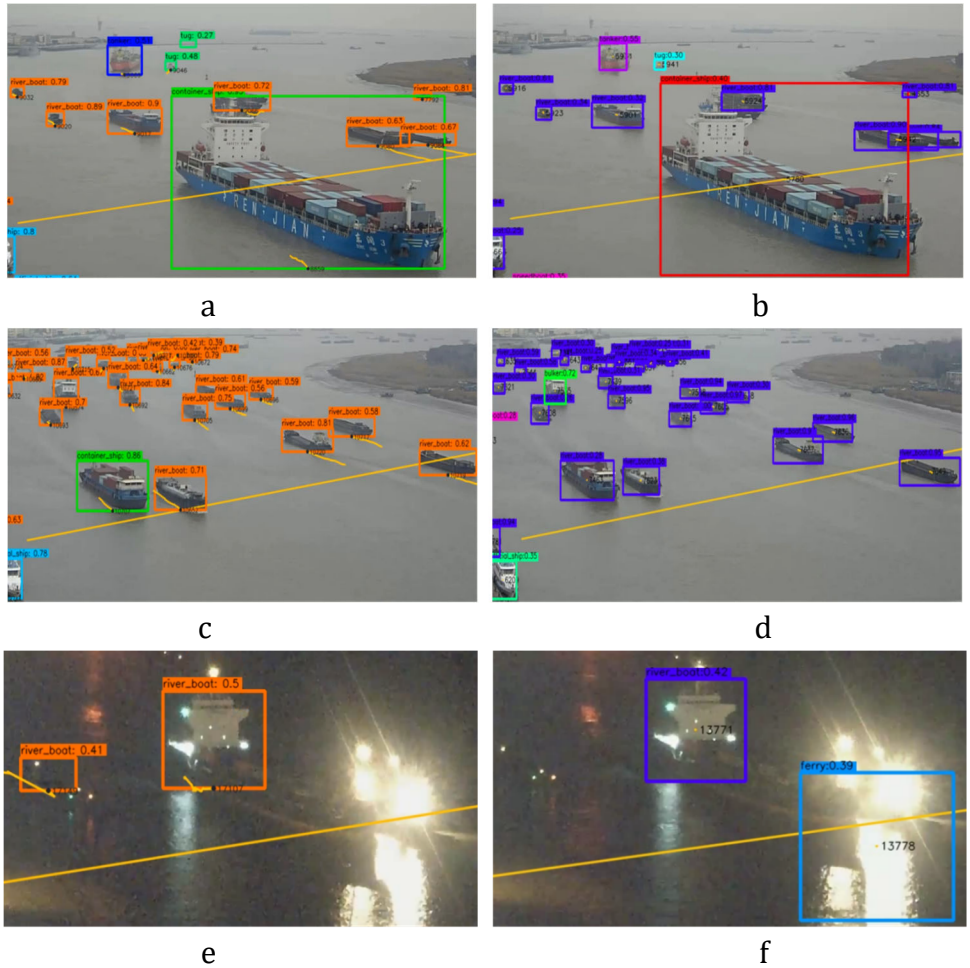


Figure 7. Joint debugging and tracking effect test chart

ISSN: 2688-8645

It can be seen from Figures (a) and (c) that CenterNet-DeepSORT can track dense ships. From the smoothness of the corresponding tracking trajectory, it can be seen that the detection frame is more stable after tracking is added, and only exists when the attitude of the ship changes greatly. Jitter. Figures (b) and (d) can be seen that some ships are not detected and classified incorrectly in YOLOv3-DeepSORT detection. The tracking algorithm deleted the tracking historical track information and reassigned the tracking identification number, which also resulted in the doubling of the tracking identification number. Figures (e) and (f) show that YOLOv3-DeepSORT has misdetection errors under strong night light conditions.

4. Summary

This article introduces target detection, multiple object tracking, SORT, and DeepSORT respectively. The coverage area of maritime surveillance is usually accompanied by complex background, illumination, deformation, and dense occlusion. This makes the pixel information (location) and category information (classification) provided by target detection often become unstable, and there are false detections and missed detections, in order to be effective To combat the above problems, the multiple object real-time tracker in this paper is designed and implemented based on the combination of CenterNet and YOLOv3 DeepSORT. After joint debugging and testing, it is found that CenterNet-DeepSORT has good real-time performance and can effectively combat the detection loss caused by detection jitter, denseness and occlusion, as well as being able to determine whether ships between different video frames are the same target, providing a basis for further research.

References

[1] Rao A, Wang H, Hu ZC, et al. A Gaussian particle filter based factorised solution to the simultaneous localization and mapping problem[C]//2013 IEEE Workshop on Advanced Robotics and its Social Impacts. IEEE, 2013: 113 -118.

[2] Yang G, Lu Q, Gao F. A novel ship detection method based on sea state analysis from optical imagery[C]//2011 Sixth International Conference on Image and Graphics. IEEE, 2011: 466-471.

[3] Xiongyong Ping, Ding Sheng, Deng Chunhua the like. Under complex weather conditions based on the depth of marine vessels detected Learning [J]. Computer Applications, 2018,38 (12): 3631-3637.

[4] Yuebang Zheng, Hansol. Improved Faster R-CNN 's SAR ship target detection method [J]. Computer and Modernization, 2019 (09): 90-95 + 101.

[5] Liu Bo, Sheng n, Zhaojian Sen, et. Based Darknet network and YOLOv3 vessel tracking recognition algorithm [J]. Computer Applications, 2019 (2019 years 06): 1663-1668.

[6] Li Y, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration[C]. European conference on computer vision. Springer, Cham, 2014: 254-265.

[7] Chen D, Yuan Z, Wu Y, et al. Constructing adaptive complex cells for robust visual tracking[C]. International Conference on Computer Vision. 2013: 1113-1120.

[8] Zhou Yong . Research and implementation of inland watercraft intelligent monitoring system based on computer vision technology [D]. Shanghai Jiaotong University, 2016.

[9] Chen Z, Li B, Tian LF, et al. Automatic detection and tracking of ship based on mean shift in corrected video sequences[C]. International Conference on Image Vision and Computing, 2017: 449-453.

[10] Grabner H, Leistner C, Bischof H, et al. Semi-supervised On-Line Boosting for Robust Tracking[C]. European conference on computer vision, 2008: 234-247.

[11] Zhang P, Wang K, Zhang L, et al. An Evolutionary Vulnerability Detection Method for HFSWR Ship Tracking Algorithm[C]. Simulated evolution and learning, 2017: 763-773.

[12] Wu G, Lu W, Gao G, et al. Regional deep learning model for visual tracking[J]. Neurocomputing, 2016: 310-323.

ISSN: 2688-8645

[13] Nam H, Han B. Learning Multi-domain Convolutional Neural Networks for Visual Tracking[C]. computer vision and pattern recognition, 2016: 4293-4302.

[14] Danelljan M, Bhat G, Khan FS, et al. ECO: Efficient Convolution Operators for Tracking[C]. Computer vision and pattern recognition, 2017: 6931-6939.

[15] Wei Wei, Yang Ru, Zhu Ye. Improved CenterNet-based remote sensing image target detection [J/OL]. Computer Engineering and Applications: 1-10 [2021-01-1]

[16] Zhou X, Wang D, Krahenbuhl P, et al. Objects as Points[J]. arXiv: Computer Vision and Pattern Recognition, 2019.

[17] Law H, Deng J. CornerNet: Detecting Objects as Paired Keypoints[C]. european conference on computer vision, 2018: 765-781.

[18] Li Xi, Cha Yufei, Zhang Tianzhu, et al. Overview of target tracking algorithms for deep learning [J]. Journal of Image and Graphics, 2019, 24(12): 2057-2080.

[19] Henriques JF, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[20] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]. international conference on image processing, 2016: 3464-3468.

[21] Wojke N, Bewley A, Paulus D, et al. Simple online and realtime tracking with a deep association metric[C]. international conference on image processing, 2017: 3645-3649.