Depth tracking of occluded ships based on SIFT feature matching

Yadong Liu[1], Yuesheng Liu[2], Ziyang Zhong[2], Yang Chen[3], Jinfeng Xia[3], Yunjie Chen[1,4,5*]
[1]School of Mathematics and Statistics,Nanjing University of Information Science, Nanjing, 210044, China
[2]ShenZhen Martime Safety Administration of China, ShenZhen, 518032, China
[3]CSIC Pengli (Nanjing) Atmospheric Ocean Information System Co., Ltd, Nanjing, 211106, China
[4]Center for Applied Mathematics of Jiangsu Province, Nanjing University of Information Science and Technology, Nanjing 210044, China
[5]Jiangsu International Joint Laboratory on System Modeling and Data Analysis, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[e-mail : 001413@nuist.edu.cn, priestcyj@nuist.edu.cn]
*Corresponding author: Yunjie Chen

Received November 8, 2022; revised February 28, 2023; accepted March 20, 2023; published April 30, 2023

Abstract

Multi-target tracking based on the detector is a very hot and important research topic in target tracking. It mainly includes two closely related processes, namely target detection and target tracking. Where target detection is responsible for detecting the exact position of the target, while target tracking monitors the temporal and spatial changes of the target. With the improvement of the detector, the tracking performance has reached a new level. The problem that always exists in the research of target tracking is the problem that occurs again after the target is occluded during tracking. Based on this question, this paper proposes a DeepSORT model based on SIFT features to improve ship tracking. Unlike previous feature extraction networks, SIFT algorithm does not require the characteristics of pre-training learning objectives and can be used in ship tracking quickly. At the same time, we improve and test the matching method of our model to find a balance between tracking accuracy and tracking speed. Experiments show that the model can get more ideal results.

Keywords: Multi-target tracking, DeepSORT, SIFT, Feature Points, Deep Learning

1. Introduction

With the progress of science and technology, target tracking technology is used more and more frequently in practical problems, such as vehicle autopilot, unmanned aerial vehicle tracking, ship tracking, and so on. Ship tracking provides important on-site micro dynamic traffic information, which is conducive to the comparative analysis of ship information entry, maritime traffic flow analysis, and ship safety improvement because it has the advantages of reducing maritime traffic risks and improving maritime traffic efficiency.

Target tracking is a very important research topic. It uses the up and down frame information of the video sequence to analyze and model the appearance and motion information of the target to predict and calibrate the target location. Target tracking combines theory and algorithms from many domains, such as depth learning, image processing, machine learning, and so on. It also paves the way for higher-level image tasks. In the early stage of target tracking research, its main research focuses on the optical flow method [1], mean shift [2,3], Kalman filter [4], particle filter [5], and other methods [6].

The concept of optical flow was first proposed by Gibson in 1950. It refers to the movement of a target, scene, or camera when moving between two consecutive frames of images. In 1994, Shi and Tomasi proposed KLT (Kanade Lucas Tomasi) optical flow method [1] for tracking. Its main idea is to use the position change of feature points between two consecutive frames to obtain the target tracking results. The Mean shift [2] algorithm was proposed by Fukunaga in 1975. Comaniciu et al. [3] used the mean shift algorithm for tracking, which iteratively solved the local maximum of the density function through gradient rise. The mean shift algorithm was fast at that time, and it was robust to target deformation and occlusion. The mean shift algorithm was widely valued. However, the extracted color histogram features have a limited ability to describe the target and lack spatial information, so the mean shift algorithm can only be used when the target and background can be distinguished in color, which has greater limitations. The tracking method based on the Kalman filter [4] predicts the target position in the next frame through state equation and historical observation data, but it is only applicable to linear systems. Because target tracking often involves nonlinear problems, particle filter algorithm which can solve nonlinear and nonGaussian filtering problems is very popular. A particle filter uses a group of random samples to approximate the target distribution and estimates the target state according to the weighted average of the samples. Isard et al. [5] first used a particle filter to solve the target tracking problem. Nummiaro et al. [6] used color histogram features to represent particles so that the algorithm can handle the deformation and occlusion of the target. These target tracking algorithms based on traditional methods not only have complex tracking processes, a large amount of calculation, and slow tracking speed, but also cannot adapt to dynamic changes, resulting in poor accuracy of traditional tracking algorithms. However, we cannot ignore the foundation of these traditional algorithms for this field.

So as to improve the tracking speed to a new stage, target tracking based on correlation filtering has become the focus of research. Correlation filtering can determine the similarity of two signals at a certain time. So, the tracking uses the filter for online learning and then calculates the correlation between the candidate area of the task object and the filter. The same signal often has the highest correlation and maximum response value. At this time, the position with the highest output response value is selected as the prediction position of the current frame of the tracking object. And then extract its feature information again to achieve the

feature update of the filter, to achieve continuous tracking of subsequent frames.

In 2010, Bolme et al. [7] took the lead in combining correlation filtering with target tracking and proposed the Minimum Output Sum of Square Error (MOSSE) algorithm, which uses the Fast Fourier Transform (FFT) to calculate in the frequency domain. With its high tracking speed and good tracking performance, it makes target tracking rise to a new stage. In 2014, Henriques et al. [8] further improved the correlation filtering framework based on MOSSE and proposed the Kernelized Correlation Filters (KCF) method. This method uses a ridge regression model to explain the correlation filtering model, adds regularization, and introduces the kernel method. The KCF method can also use FFT to accelerate the operation, which can achieve higher accuracy while maintaining a high frame rate. After KCF, many methods based on correlation filtering have been proposed. These methods expand the correlation filtering methods and improve the performance. The main methods include introducing multi-scale, adding multiple features, block tracking, weakening boundary effects, and designing better and more complex loss functions to train better filters. For better tracking performance of correlation filtering, multiple features were introduced, including Histogram of Orientated Gradient (HOG) features [9], Color Names (CN) features [10], convolution features, etc.

For tracking occluded objects, some block-tracking methods are proposed. Li et al. [11] used particle filtering to sample tracking blocks and position prediction and fused the tracking results by weighting. The weight is equal to the weight of the particles. Liu et al. [12] used fixed target blocks to evaluate the tracking results to obtain weighted weights. The target itself has a spatial structure. Simply tracking multiple blocks separately does not consider the relationship between multiple blocks. Liu et al. [13] assumed that the motion of each block is similar, defined that the motion vector of each block is equal to a common motion vector plus a sparse term, and simultaneously solved the filter coefficients of each block by minimizing the sparse term. Block tracking can deal with partial occlusion to some extent because some targets are visible under partial occlusion, and visible target blocks can still be tracked, but targets under full occlusion cannot be tracked. Although these algorithms based on correlation filtering accelerate the tracking speed and improve the tracking accuracy compared with the traditional methods, they still have great obstacles in practical application, because the correlation filtering algorithm requires high computing power in computing, the model structure is complex, and there are certain limitations in occlusion.

With the improvement of computer equipment performance and the hot trend of deep learning, researchers will apply the advantages of deep learning to multi-target tracking to achieve the integration of deep learning and target tracking. Now the multi-target tracking methods based on depth learning include DFT and DBT according to the different initialization methods. DFT algorithm uses the first frame detection frame to continuously track the motion of the detection frame to generate the target moving track. This method can improve the tracking speed by reducing the amount of computation, but its shortcomings are also very obvious. First, it is impossible to track new targets in subsequent frames; Secondly, if the previously selected target disappears from the image, the tracking information will be lost. So the tracking accuracy of this tracking method is poor. Detection-based multi-target tracking has become the mainstream of multi-target tracking research, and detection-based multi-target tracking includes online tracking and offline tracking due to different processing methods. Offline tracking can overcome the limitations of data association in a limited time, and improve the performance of multi-target tracking in complex situations such as target occlusion, detector missing, etc. It has a broad application space in offline video retrieval and analysis. However, the disadvantage of this batch frame processing method is that it increases

the computational complexity, and the global optimization also involves future information. However, most of the actual tracking scenes cannot submit predicted future information, so this method cannot accomplish the task of real-time tracking.

Since the fusion of deep learning and multi-target tracking, the multi-target tracking method based on online data association of detection has attracted more and more attention. In earlier studies, in order to accurately determine the target trajectory, target detection response and filtering technology were combined to simply and directly implement the tracking using detection. For example, some methods are combined with the Kalman filter. In the process of multi-target tracking, the Kalman filter is tried to solve the problem of target position prediction and estimation, which can help to screen the correct target detection response as an associated target. Wang et al. [14] used the inter-frame difference method to detect moving targets, adaptively marked the label of each target, assigned an independent Kalman filter to each target to predict the target position, and achieved data association via the Hungarian algorithm. Eiselein et al. [15] proposed a multi-target tracking method based on Gaussian Mixture Probability Hypothesis Density (GMPHD) and fused the multi-target detection results to achieve accurate multi-target tracking. Bae and Yoon [16] proposed a multi-target tracking method based on trajectory confidence coefficient and online discriminant apparent learning, established the target trajectory segment and trajectory confidence coefficient, according to the different values of the confidence coefficient, associated the target trajectory segment with the detection response, or the target trajectory segment with other trajectory segments in hierarchical data, which can effectively reduce the tracking error caused by complex situations such as background changes, frequent occlusion, etc. However, the speed and accuracy of these algorithms are not improved much. Until 2016, the SORT algorithm was proposed by Alex Bewley et al. [17] But the disadvantages are also obvious. Because SORT only uses the Kalman filter to estimate the motion state of the target and ignores the similarity of the target itself, causing frequent identity switching. The scene that will be occluded cannot be tracked effectively. The next year, DeepSORT [18] algorithm was proposed, which added a feature extraction module and cascade matching method on the basis of SORT to measure the similarity of targets. DeepSORT improves the accuracy of target tracking again without reducing the tracking speed. At the same time, it can also achieve effective tracking in complex occlusion scenes. The DeepSORT algorithm is widely used because of its high tracking speed, high precision, and simple model. However, its feature extraction model is not suitable for all scenes, especially for objects whose features are not obvious, such as the ship target in our study, under the influence of lighting and other factors, it is likely causing the feature extraction model to incorrectly identify the target. Also, the feature extraction model needs to be trained in advance, so it cannot be applied to some specific scenes in real-time. In terms of occlusion, the DeepSORT algorithm is proposed as a short occlusion tracking algorithm for pedestrian tracking, which has some limitations when applied to long occlusion problems such as ship tracking.

Therefore, this paper gives a DeepSORT algorithm based on SIFT features, which overcomes the shortcomings of low tracking accuracy of traditional depth learning models in the case of ships with or without occlusion by using the advantages of SIFT features such as scale invariance, illumination invariance, etc.

2. Related Work

2.1 Object Detection

As the model input of detection-based multi-target tracking, target detection is very important in the tracking process. The detection effect of the detector often directly affects the tracking accuracy. Nowadays, target detection is developing rapidly, and there are many good detection algorithms that can be used in tracking tasks. If you want to apply a detection algorithm to target tracking, you must have high precision and high speed to meet the requirements of real-time target tracking. At present, the detection algorithms that are very popular and highly practical belong to the YOLO series, especially YOLOv5. YOLO detection algorithm belongs to one-stage target detection. It has a faster detection speed because it does not require the region proposal phase but instead generates the class probability and location information of the object directly. YOLOv1 [19] is the first one-stage of the deep learning detection algorithm. Its detection speed is very fast. The algorithm divides the image into multiple grids, and then predicts the boundary box of each grid at the same time and gives the corresponding probability. As it is the early generation of the YOLO series, it only considers the detection speed and does not enhance the detection accuracy, especially for small targets.

Based on YOLOv1, YOLOv2 [20] has made great improvements in accuracy, speed, and classification quantity. In terms of speed, YOLOv2 uses DarkNet19 [20] as the feature extraction network. In terms of classification, YOLOv2 uses the joint training skills of target classification and detection, combined with Word Tree and other methods, to expand the detection categories of YOLOv2 to thousands. In terms of accuracy, YOLOv2 uses a series of methods such as Batch Normalization to greatly improve the detection accuracy. However, the YOLOv2 algorithm has only one detection branch, and the network lacks the capture of multi-scale context information, so the detection effect for targets of different sizes is still poor, especially for small target detection. Compared with YOLOv2, YOLOv3 [21] replaced the feature extraction network with DarkNet53 [21], replaced Softmax with Logistic for object classification, and used the FPN [22] idea to detect objects with different sizes by using three branches. YOLOv4 [23] introduced Mosaic data enhancement, cmBN, and SAT self-confrontation training at the input end; On the feature extraction network, YOLOv4 uses various good methods, including CSPMarknet53 [23], Mish activation function [24], Dropblock [25]; In the detection head, SPP module [26] is introduced and FPN+PAN [27] structure is used for reference; In the prediction stage, CIOU [28] replaces the always used IOU, and NMS is replaced by DIOU_ NMS, etc. YOLOv4 introduced the latest research in deep learning in previous years to YOLOv4 for verification testing, which is a big step forward on the basis of YOLOv3. YOLOv5 is similar to YOLOv4 in that it also integrates a large number of the latest technologies, thus significantly improving the detection performance. Although YOLOv5 is not as good as YOLOv4 in performance, it performs better in speed and flexibility and is widely used because of its superiority in the rapid deployment model. Therefore, this paper uses the YOLOv5 model to obtain ship position information.

2.2 Object Tracking

SORT is a multi-target tracking algorithm that can effectively associate targets and improve tracking in real-time in recent years. Its main core is to use the Kalman filter to predict the position of the target frame by frame to match the IOU of the detected target, and finally, use the Hungarian algorithm to achieve optimal pairing. This simple method achieves good performance at high frame rates: fast speed and low computational resource consumption. At that time, the tracking speed reached 260HZ, which was 20 times faster than other methods.

But its disadvantage is also obvious: sort has a strong dependence on detection. When the target is not detected or is occluded, the sort will directly determine that the target disappears. Even if the target is detected again in a later frame, it will be considered another target. Now the improvement of target detection performance has solved the most basic input problem of target tracking, making the target tracking algorithm not frequently switch IDs without occlusion, but occlusion is still the main problem faced by target tracking.

The DeepSORT algorithm is an improvement and optimization of the existing problems of the SORT algorithm. It solves the problems of frequent switching of target labels and the inability to effectively track under occlusion. DeepSORT algorithm continues to use the core idea of SORT, but on this basis, they also use a feature extraction model to extract and store target features. After that, the tracking accuracy is greatly improved by feature matching and cascade matching methods, and it can be effectively tracked in the occluded scene. Now DeepSORT algorithm combined with an efficient YOLOv5 detector is a very popular detection and tracking algorithm, which has a very good effect in pedestrian tracking, vehicle tracking, and other applications.

3. Methods

3.1 Track Prediction and State Determination

Kalman filter is used to predict the next frame trajectory state information of the target. We use the eight-dimensional state space $(\mathrm{u}, \mathrm{v}, \gamma, \mathrm{h}, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$ to describe, They correspond to the frame center coordinate, aspect ratio, height and respective speed. Kalman predicts the movement information of the target:

$$x' = Fx \tag{1}$$

$$P' = FPF^T + Q \tag{2}$$

where x is the mean value of the previous frame and x' is the predicted value of the current frame, F is the state transition matrix, P is the covariance of the previous frame and P' is the predicted value the current frame, Q is the noise matrix. Kalman updates the movement information of the target:

$$y = z - Hx' \tag{3}$$

$$K = P'H^T(HP'H^T + R)^{-1} \tag{4}$$

$$x = x' + Ky \tag{5}$$

$$P = (I - KH)P' \tag{6}$$

z is the mean vector of the detection, excluding the velocity change value, H is responsible for mapping the mean vector x' to the detection space and y is the mean error of the detection and trajectory. R is the matrix of the detector, which includes the center point and wide and high noise. K is used to estimate the importance of the error, x and P are updated mean vector and covariance matrix respectively.

The following rules govern the determination of a target track: 1. The track is designated as a Tentative state at first; 2. It will be changed to confirmation status after three frames are matched successfully; 3. If the track in the temporary state is not matched in the next frame or if the track in the confirmation state is not matched during a set frame length, it is determined to be in the deletion state for the convenience of subsequent deletion of the track.

3.2 Extract feature points using SIFT algorithm

SIFT feature [29] is invariant to rotation, scaling, and brightness changes, and also adaptable to changes in view angle, affine transformation, and noise. In addition, SIFT operator has the characteristics of multiplicity and high speed. A few objects can generate a major number of SIFT feature vectors, so it is widely used for target matching. The order of extracting feature points: 1. Scale space extremum detection: searching image positions on all scales. Gauss differentiation is used to identify points that are invariant to scaling and rotation. 2. Key point positioning: determine the position and proportion through the fine-fitting model to select stable key points on candidate positions. 3. Determination of key point direction: since the subsequent operations on the image are based on the direction, scale, and position of key points, we need to specify one or more directions for each key point position according to the local gradient direction of the image to provide invariance for these transformations. 4. Key point descriptor generation: in neighborhoods near key points, the local gradient of the image is measured based on the scale to allow larger local shape distortion and illumination changes. As shown in **Fig. 1**, the number of similar key points can be determined with the key points of the target to achieve feature matching.

Fig. 1. Ship feature points extracted by SIFT algorithm

3.3 Target Matching

Our method is still based on IOU matching to reduce the probability of error matching and missing matching after feature matching. At the same time, we introduce SIFT feature matching to improve the model tracking accuracy. After the traditional SIFT algorithm extracts the feature points of the prediction target and the detection target, it usually uses BF Matcher or Flann-Based Matcher to calculate the number of similar feature points between targets. BF Matcher will try all possible matches so that it can find the best match. For many feature points, the speed of this matching method is slow. Flann is an approximation method, which is faster but finds the nearest neighbor approximation. When there are many target feature points, so as to ensure real-time performance, we use Flann-Based Matcher for matching. When there are not many feature points, we directly use BF Matcher to calculate the number of similar points to ensure the accuracy of matching. **Fig. 2** is the results of our research goal under two calculation methods. It can be seen that BF Matcher meets the requirements of ship tracking.

After obtaining the results of similar points between multiple prediction and detection targets, we use the matching ratio P to express the similarity in order to make it easier for the computer to achieve accurate matching between the targets better and faster.

$$P = \frac{M}{max(p1,p2)} \tag{7}$$

where $p1$ is the feature point extracted from the target of the detection box, $p2$ is the feature point of the target in the prediction box, M is the statistics of similar feature points of the two targets. The larger P, the better the matching result. That is, we can think that the target is the same target, but when the matching rate is maximum and does not exceed the threshold, we think that the matching fails and there is no target to match with.

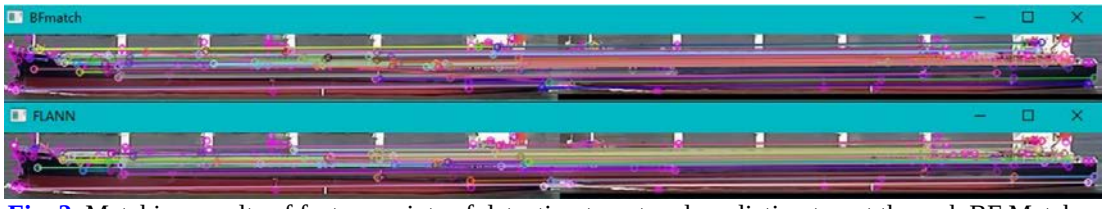

Fig. 2. Matching results of feature points of detection target and prediction target through BF Matcher method and Flann Based Matcher method.

3.4 Depth Match

Fig. 3 is the flowchart of our algorithm. Depth matching includes cascade matching and multi-frame feature matching. Using depth matching is to enable the algorithm to track effectively in the occluded scene, and further enhance the tracking accuracy. Occlusion is always the biggest problem in target tracking. For this question, we use a depth-matching model based on SIFT features to solve it. Different from the traditional tracking idea, this paper no longer regards the whole problem as a global optimization problem but uses a cascade strategy to optimize a series of subproblems. If a long occlusion occurs, the uncertainty related to the target position predicted by the Kalman filter will increase greatly. If the detected target matches multiple tracks at this time, the track that disappears longer tends to be more uncertain in tracking and predicting locations, that is., the covariance is greater. The inverse of the covariance is used in the calculation of Markov distance, so the detection target is more inclined to match the track that disappears longer. This undesirable effect often causes discontinuity in tracking. Therefore, cascade matching gives more weight to frequently occurring targets, that is, whenever matching, the trajectory with the same occlusion time is considered.

We store the feature point information extracted when the target is detected in multiple frames for a long time, that is, the model has multiple feature point information extracted from each target in different frames. In feature matching, we match the feature points of the target's current frame with those of all frames stored in all tracks and calculate the best result as the final result. Multi-frame feature matching is to prevent the unexpected situation (attitude change or incomplete target detection) that occurs when the target extracts feature points before occlusion.

The Hungarian algorithm is used to achieve optimal pairing of detection boxes and tracks. We extract feature points of all detection frame targets in the current frame and construct the feature matrix x_j, and all the feature information of the N frame of the track is stored in X_i, then the feature matrix for each frame of the track is x_i. Among them, j is the target of the detection box, i represents the target of the prediction box, and N is a value that requires us to find to improve the tracking accuracy while also achieving real-time tracking. The minimum distance between the ith target track feature and the jth target detection feature is:

$$d_{(i,j)} = min\left\{1 - x_j^T x_i^{(k)} \mid x_i^{(k)} \in X_i, k \in (1, N)\right\} \tag{8}$$

A control threshold t is introduced to the eigenvalue distance to determine whether the two are allowed to be correlated, which is calculated by the following formula:

$$P_{(i,j)} = \begin{cases} 1, d_{(i,j)} \leq t \\ 0, d_{(i,j)} \geq t \end{cases} \tag{9}$$

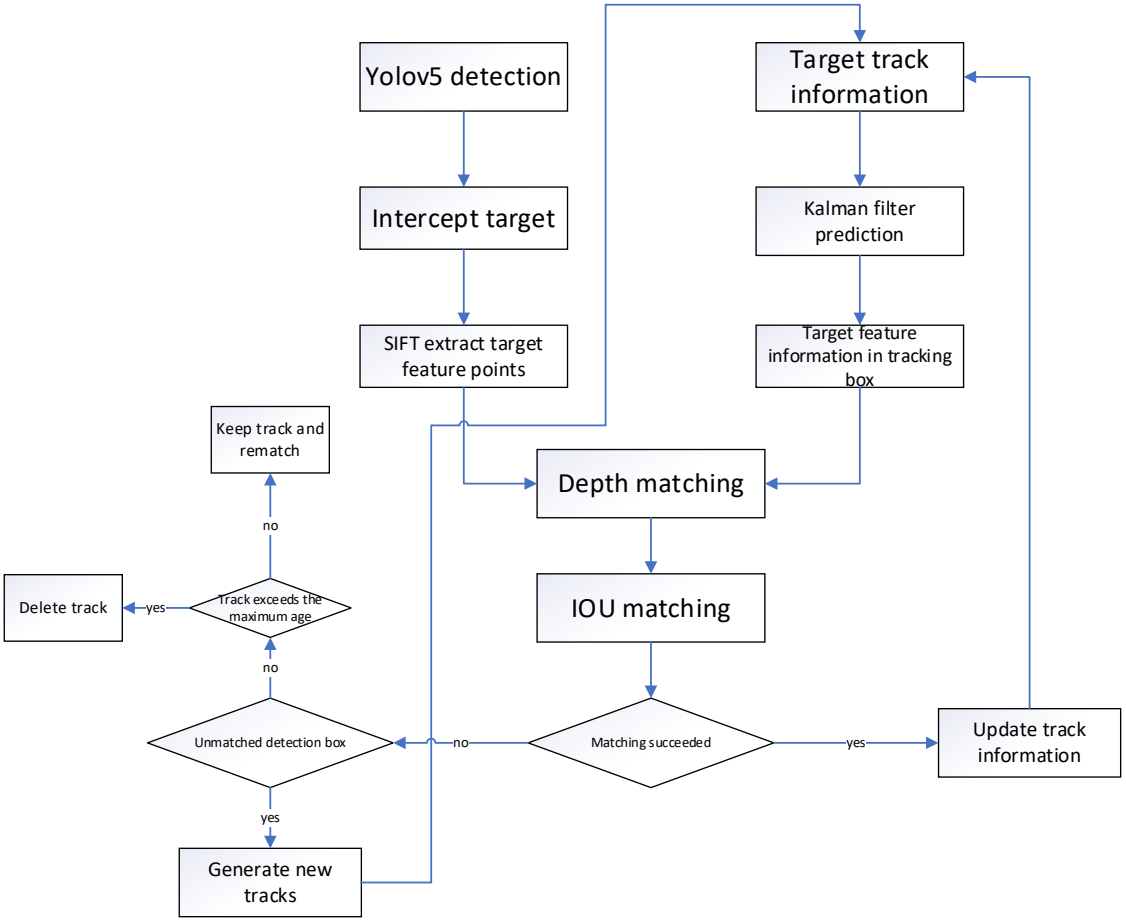

Fig. 3. Our algorithm flowchart.

4. Experiments

4.1 Datasets Description

The data in this paper is a real-time channel acquisition data set, including 1000 segments of ship video data with a duration of more than one minute. It includes all the scenes required for tracking experiments: 1. Tracking when a ship first appears on the screen; 2. Normal and unobstructed ship tracking; 3. Tracking of target reappearance under long-term occlusion caused by overtaking ship; 4. The tracked ship disappears in the picture. We selected 40 videos as datasets, including 20 videos with ship occlusion and 20 videos without ship occlusion. Each video captured about 200 pictures totaling 10000 pictures as training datasets.

4.2 Parameter Selection and Evaluation Indicators

Because ships are generally long, when overtaking ships, the blocked ships often disappear for tens of seconds or even longer, so this model requires that target information can be retained longer for successful matching when subsequent targets appear again. In this paper, 2000 frames are used to evaluate the test sequence, and the detection threshold is 0.3. The only part of our method that needs network training is the YOLO detection part. Because the focus of our attention is not on the detection process but on the tracking process, we have not changed too many parameters in the ship detection training. We used 16 batch sizes to train 300 epochs, the optimizer was ADAM optimizer, and the model learning rate was set to the learning rate decay mode.

MOTA [30] and IDF1 [31] are often considered the most important evaluation indexes when testing the performance evaluation of tracking algorithms on MOT data. The indicators related to MOTA include false negatives, false positives, and ID switch rate, where false negatives and false positives are detection results, and only ID switch is related to tracking. MOTA is more concerned with the performance of the detector. Unlike MOTA, IDF1 focuses more on relevance and consistency and mainly calculates F1 values with correct IDs on a track. When comparing tracking algorithms with the same or smaller ID switches, the IDF1 indicator is more representative. The following are the evaluation indicators:

- IDF1(↑): The proportion of detected objects with correct ID among detected and tracked objects.

$$IDF1 = \frac{2IDTP}{2IDTP+IDFP+IDFN} \tag{10}$$

  IDTP can be seen as detecting the number of correctly allocated targets throughout the video, IDFN can detect the number of missed allocations of targets throughout the video, and IDFP can detect the number of incorrectly allocated targets throughout the video.
- IDP(↑): Identification precision.

$$IDP = \frac{IDTP}{IDTP+IDFP} \tag{11}$$

  IDR(↑): Identification recall.

$$IDR = \frac{IDTP}{IDTP+IDFN} \tag{12}$$

- ID Sw(↓): The total number of identity switches.
- MOTA(↑): This measure combines three error sources: false positives, missed targets, and identity switches.

$$MOTA = 1 - \frac{\sum_t (FN_t+FP_t+IDSW_t)}{\sum_t GT_t} \tag{13}$$

4.3 Results and Analysis

Table 1 shows the average target tracking accuracy of the model in this paper. Compared with the original DeepSORT algorithm, our method improves by 2.8% on IDF1, 3.2% on IDP, and 2.6% on IDR. It shows that our method maintains track consistency throughout the tracking process. **Fig. 4** is the DeepSORT model tracking results and the algorithm tracking results in this paper. We can see that in the whole tracking process, the method in this paper does not

have an incorrect ID switch, while the DeepSORT algorithm incorrectly identifies the ship when the occluded ship reappears, leading to the ID switch, and only a few seconds later did it correctly identify the ship ID. Combined with **Table 1**, the reduction of ID Sw shows that our method has a low error rate for the identification of the same ship.

In this experiment, we test our method on an N-fit selection experiment. Because the gap between frames is small, instead of taking the last consecutive frame before the target disappears as a sample, we took a Frame-breaking sample and stored the feature every 100 frames (about 3 seconds). From **Table 2**, we can see that tracking occlusion targets fails when N = 1. We only select target-frame feature information to store matches. Frames that are most likely stored are detected as incomplete targets. When subsequent targets have detected features and this frame information matches beyond the threshold, their associations are considered not allowed. When N = 3 and 5, the obscured target can be tracked effectively, and the tracking speed can meet the real-time requirements. With the increase of N, the model is more robust, but the corresponding tracking speed will decrease. N=5 is the best value we recommend.

Table 1. Tracking results on our ship data.

	IDF1↑	IDP↑	IDR↑	IDSW↓	MOTA↑	FP↓	FN↓
SORT	81.7%	92.5%	73.1%	2	75.7%	87	1182
Deep SORT	83.7%	94.7%	74.9%	2	75.7%	88	1180
Ours	86.5%	97.9%	77.5%	0	75.8%	87	1178

Fig. 4. DeepSORT algorithm and Our algorithm for ship tracking.

Table 2. Test of N value.

N(frame)	IDF1(%)	FPS
1	81.7	41
3	86.5	35
5	86.5	31
10	86.5	25
20	86.5	17

5. Conclusion

This paper proposes a deep sort model based on SIFT features to reduce the impact of occlusion on target tracking accuracy. The model makes use of the scale invariance and

illumination invariance of SIFT operator to reduce the influence of ship size change and attitude change after occlusion. At the same time, SIFT feature information in the time series is considered to reduce the impact of poor target trajectory matching accuracy under long-term occlusion. How to use the latest technology to improve tracking accuracy and speed is the focus of our follow-up research.

Acknowledgement

This work was supported in part by the Six Talent Peaks Project in Jiangsu Province SWYY-034, Natural Science Foundation of Jiangsu Province of China BK20191394 and the National Nature Science Foundation of China 61672291.

References

[1] Jianbo Shi and Tomasi, "Good features to track," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593-600, 1994. Article (CrossRef Link)
[2] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32-40, January 1975. Article (CrossRef Link)
[3] D. Comaniciu, V. Ramesh and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 142-149, 2000. Article (CrossRef Link)
[4] Kalman, Rudolph Emil, "A new approach to linear filtering and prediction problems," *ASME.J. Basic Eng*, vol. 82, no. 1, pp. 35-45, March 1960. Article (CrossRef Link)
[5] Isard, Michael, and Andrew Blake, "Condensation—conditional density propagation for visual tracking," *International journal of computer vision*, vol. 29, no. 1, pp. 5-28, 1998. Article (CrossRef Link)
[6] Nummiaro, Katja, Esther Koller-Meier, and Luc Van Gool, "An adaptive color-based particle filter," *Image and vision computing*, vol. 21, no. 1, pp. 99-110, 2003. Article (CrossRef Link)
[7] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2544-2550, 2010. Article (CrossRef Link)
[8] J. F. Henriques, R. Caseiro, P. Martins and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583-596, 1 March 2015. Article (CrossRef Link)
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886-893, 2005. Article (CrossRef Link)
[10] Danelljan, Martin, et al, "Adaptive color attributes for real-time visual tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1090-1097, 2014. Article (CrossRef Link)
[11] Li, Yang, Jianke Zhu, and Steven CH Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 353-361, 2015. Article (CrossRef Link)
[12] Liu, Ting, Gang Wang, and Qingxiong Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4902-4912, 2015. Article (CrossRef Link)
[13] Liu, Si, et al, "Structural correlation filter for robust visual tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4312-4320, 2016. Article (CrossRef Link)

[14] Wang, Zhenhai, et al, "An online multi-object tracking approach by adaptive labeling and kalman filter," in *Proc. of the 2015 Conference on research in adaptive and convergent systems*, pp. 146-151, 2015. Article (CrossRef Link)
[15] V. Eiselein, D. Arp, M. Pätzold and T. Sikora, "Real-Time Multi-human Tracking Using a Probability Hypothesis Density Filter and Multiple Detectors," in *Proc. of 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pp. 325-330, 2012. Article (CrossRef Link)
[16] Bae, Seung-Hwan, and Kuk-Jin Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1218-1225, 2014. Article (CrossRef Link)
[17] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. of 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464-3468, 2016. Article (CrossRef Link)
[18] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. of 2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645-3649, 2017. Article (CrossRef Link)
[19] Redmon, Joseph, et al, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016. Article (CrossRef Link)
[20] Redmon, Joseph, and Ali Farhadi, "YOLO9000: better, faster, stronger," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6517-6525, 2017. Article (CrossRef Link)
[21] Redmon, Joseph, and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv: 1804. 02767*, 2018. Article (CrossRef Link)
[22] Lin, Tsung-Yi, et al, "Feature pyramid networks for object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936-944, 2017. Article (CrossRef Link)
[23] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv: 2004. 10934*, 2020. Article (CrossRef Link)
[24] Misra, Diganta, "Mish: A self-regularized non-monotonic neural activation function," *arXiv preprint arXiv: 1908. 08681*, 4.2, 10-48550, 2019. Article (CrossRef Link)
[25] Ghiasi, Golnaz, Tsung-Yi Lin, and Quoc V. Le, "Dropblock: A regularization method for convolutional networks," *Advances in neural information processing systems*, 31, 2018.
[26] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 1 Sept 2015. Article (CrossRef Link)
[27] Wang, Wenhai, et al, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8439-8448, 2019. Article (CrossRef Link)
[28] Zheng, Zhaohui, et al, "Distance-IOU loss: Faster and better learning for bounding box regression," in *Proc. of the AAAI conference on artificial intelligence*, vol. 34, no. 07, pp. 12993-13000, 2020. Article (CrossRef Link)
[29] Lowe, David G, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 02, pp. 91-110, 2004. Article (CrossRef Link)
[30] Bernardin, Keni, and Rainer Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, pp.1-10, 2008. Article (CrossRef Link)
[31] Ristani, Ergys, et al, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. of European conference on computer vision*, Springer, Cham, pp. 17-35, 2016. Article (CrossRef Link)

Yadong Liu received a bachelor's degree in statistics from Xuzhou institute of technology, Xuzhou, China, in June 2021.He is currently pursuing the M.S. degree with the School of Mathematics and Statistics, Nanjing University of Information &Science, Nanjing, China. His research interests include image processing and deep learning.

Yuesheng Liu received the Master degree in electronic engineering from Shanghai Maritime University, Shanghai, China in 1999. He is the Deputy Director of Scientific and Technical Information Division of Shenzhen Maritime Safety Administration. His research interest is mainly focused on pattern recognition, Ship dynamic monitoring and ship identification.

Ziyang Zhong received the Master degree in Maritime Affairs- Maritime Safety and Environmental Management from World Maritime University, Malmo, Sweden in 2016. He is a chief staff member of Scientific and Technical Information Division of Shenzhen Maritime Safety Administration. His research interest is mainly focused on telecommunication and Intelligent ship monitoring.

Yang Chen received the Ph.D. degree in mechanical engineering from Nanjing University of Science & Technology, Nanjing, China, in 2010. His research interest is mainly focused on the overall technology research of shore based water surface surveillance radar, the overall technology research of shipborne navigation radar, and the overall technology research of ship shore cooperation.

Jinfeng Xia recesiced the Master degree in traffic information engineering and control from Dalian Maritime University in 2005. At present, he is a professional in VTS product development of CSIC Prade (Nanjing) Atmosphere and Information System Co., Ltd. His research interests mainly focus on tracking algorithms and image processing.

Yunjie Chen received the Ph.D. degree in Pattern recognition and intelligent system from Nanjing University of Science and Technology, Nanjing, China, in 2008. He is currently a professor with the School of Math and Statistics, Nanjing University of Information Science and Technology, Nanjing. His research interest is mainly focused on pattern recognition, image segmentation, and image processing.