Ranging Research on Telematics Based on Mask R-CNN Dual Eye Stereo Vision Ranging Algorithm

Shuangshuang Chen*[1], Xiaojie Li, Kai Wang, Jibin Sun, Bo Yang
Yantai Automobile Engineering Professional College, China

ABSTRACT

With the increasing popularity of automobiles, the occurrence of safety accidents is on the rise. Under the development of deep learning, the car networking system has been developed rapidly, and automobile intelligence has become possible. In order to meet the demand for body environment detection of intelligent vehicles and improve their detection performance, a large number of existing related target detection studies have been conducted, but there is still room for improvement in terms of accuracy and efficiency. The study proposes an optimized detection method of Mask Regional Convolutional Neural Network (Mask R-CNN) and combines it with SGBM binocular stereo vision ranging method to achieve target localization and distance range quantification around the car body. The results show that the method carries out the characterization of the pyramid structure (Feature Pyramid Network (FPN), MobileNet Version 3 Mobile-Segmentation (MobileNet-V3_MS), path aggregation network (PANet), and the target detection and distance range quantification. aggregation network (PANet), Channel attention (CA), Spatial attention (SA), and Semi-Global Block Matching (SGBM) optimizations have improved the performance of the Precision, recall, Dicc loss mAP and F1 are improved by 2.13%, 2.03%, 1.82%, 2.26% and 2.43% respectively. The proposed algorithm performs better compared to other image processing methods and achieves optimal performance with fewer iterations, converging to 97% accuracy at 100 iterations. Therefore the proposed method is very effective in the application of target ranging process in the environment around the car, which can improve the performance of the car networking system and provide a reference for car intelligence.

Keywords: Intelligent vehicles; Feature extraction; SGBM; Detection; Telematics

1. INTRODUCTION

The rapid development of new energy has made automobiles become consumer goods for ordinary families, bringing convenience to people's lives and becoming a common means of transportation for people traveling [1]. The increase in the number of car-using groups has led to an increase in car-using scenarios, and people have put forward new demands for the functions of cars, while the increase in the number of cars has further increased the incidence of traffic accidents [2]. With the arrival of the information age, vehicle network technology came into being, providing an important foundation for the arrival of intelligent cars, and the intelligent reform of various functions of automobiles is inevitable [3]. The reform of intelligentization of automobile functions by vehicle network technology is mainly divided into two categories: internal functions (communication and network) and external functions (environment detection, automatic driving). Among the external functions of intelligent vehicles, target detection and ranging around the body can be realized. Therefore, improving the performance of Telematics algorithms can effectively avoid the occurrence of road traffic safety accidents [4]. Currently, the common target detection methods are traditional digital techniques and algorithms based on traditional convolutional neural networks, which are unable to realize the efficient processing of a large number of image features and obtain high detection accuracy [5].

Many experts at home and abroad have also carried out a lot of research on image feature extraction. Yu B and other scholars proposed a target detection method based on PSPNet for the problem of wearing helmets at construction sites, which can solve the supervision of workers' helmet wearing by image extraction for small targets of helmets, and improve its accuracy [6]. Lu X F et al. proposed a method of target recognition and localization for the problem of crack identification, which can effectively detect and identify the targets [7]. Lab target identification and localization method, which can effectively detect and identify the target [8]. However, the above methods are unable to achieve accurate extraction of features and guarantee the detection rate when performing target detection, and do not have high accuracy

[1] *clxjxk@163.com, xiao333jie@163.com,17862837336@163.com,sunytqc917@163.com,67612707@qq.com

The International Conference Optoelectronic Information and Optical Engineering (OIOE2024), edited by
Yang Yue, Lu Leng, Proc. of SPIE Vol. 13513, 135133C · © 2025 SPIE · 0277-786X · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048771

in terms of quantifying the target distance [9]. Therefore, the study proposes an improved Mask R-CNN target detection algorithm and combines it with binocular stereo vision ranging technology based on SGBM to realize target localization and ranging in the external environment of the vehicle. The innovations of the proposed method are : Improvement of Mask R-CNN, increase the target image focus feature weights, and use of binocular stereo vision ranging method based on SGBM to improve the efficiency and accuracy of target object distance measurement.

2. BINOCULAR STEREO VISION RANGING BASED ON MASK R-CNN ALGORITHM

2.1 Improved target detection algorithm for Mask R-CNN

The development of new energy has accelerated the popularization of automobiles and brought new momentum for social and economic development. The popularization of automobiles has brought convenience and also increased the probability of traffic accidents, so the technology of Telematics has arisen. And the integration of AI and Telematics will add intelligent detection function for vehicles, thus providing protection for road traffic safety issues. Using Mask R-CNN algorithm as target detection, it can play its powerful image feature extraction function and realize target segmentation by mask branching, and its structure is shown in Figure. 1[10].

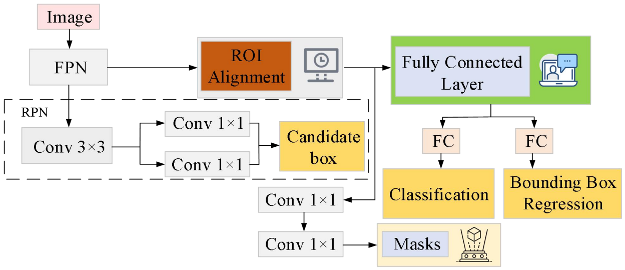


Figure 1. Mask R-CNN structure.

From the Figure 1 it can be seen that Mask R-CNN's RPN structure can realize the accurate detection of the image through the full convolution process of the image. The realization of the full convolution process is specifically divided into 3 steps, firstly, the input image is convolved to get smaller features, and then it is deconvolved to get larger features, and then each pixel of the image is predicted to be a category and distributed probabilistically through the activation function , and finally each pixel category corresponds to the appropriate Mask respectively.The RPN loss function of Mask R-CNN is shown in Eqn. (1) [11].

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{1}$$

In Eq. (1), i denotes the number of the current feature candidate area ; P_i is the predicted category label; $P_i{}^*$ denotes the real category label, which belongs to the range of [0, 1]; t_i and t_i^* denote the real and predicted category probability regression parameters, respectively; N_{cls} and N_{reg} denote the number of feature candidate area samples and number of candidate frames, respectively; $\sum_i L_{cls}(p_i, p_i^*)$ denotes the categorization loss; $\sum_i p_i^* L_{reg}(t_i, t_i^*)$ denotes the bounding box regression loss; λ denotes the $Smooth_{L1}$ function; $L(\{p_i\},\{t_i\})$ denotes the RPN loss function. FPN network can easily enhance its spatial and strong semantic information by fusing the features after convolution of multiple layers when feature extraction is performed. feature extraction, by fusing the features after multi-layer convolution to enhance its spatial and strong semantic information.The top-down transfer process adopted by FPN in the fusion process using its pyramid feature structure is prone to make it difficult for the features to reach the deeper layers smoothly from the shallow layers, thus resulting in the problem of feature information loss. Therefore, PANet is introduced so that the top-down transfer path of features is shortened, thus allowing the bottom layer of the pyramid to obtain more feature information and increase the algorithm localization capability.

In order to further enhance the fusion between feature information and improve the algorithm's ability to process image edge information, the attention mechanism is added to the RPN-FPN module, which is a CA-SA module combining CA and SA, respectively. This module increases the weights of the focus features so that the attention is focused on the target pixel points, thus enhancing the algorithm's image detection and segmentation accuracy. In addition, the two-stage computation characteristic of the Mask R-CNN algorithm makes it have a complex network structure, which guarantees its computational accuracy and is accompanied by a large number of computational parameters, resulting in an increase in the amount of computation and slowing down the computation rate. Therefore, it is lightweight and improved by proposing the feature extraction network of MobileNet-V3_MS+FPN+PANet in conjunction with MobileNet-V3_MS, which optimizes the number of convolutional kernels and channels, removes unnecessary computational parameters of the algorithm, and obtains lightweight computational steps, thus enhancing the computational rate and accuracy .

The architecture of MobileNet-V3 contributes to the lightweighting of the algorithm in two ways. On the one hand, the number of convolutional kernels in the first convolutional layer is directly halved, removing the number of redundant parameters; on the other hand, the Efficient Last Stage is used to remove the redundant steps between the convolutional layers.The lightweighting of the network architecture of MobileNet-V3 simplifies the structure of the Mask R-CNN network, and reduces the computation time, while keeping the algorithm's accuracy unchanged. The accuracy of the algorithm remains unchanged. However, there is still room to improve the accuracy of the algorithm, which can be realized through two modification steps. First, use Mish as the activation function of the network, whose gradient smoothness can achieve the optimization of the results; the non-monotonicity of the Mish function can achieve the retention of the minimum value of the function and prevent the gradient from disappearing. Second, the channel features are weighted using an excitation operation to emphasize the focal features of the shallow convolutional layer.

2.2 SGBM-based binocular stereo vision ranging technique

In order to protect road traffic safety, vehicles need to be equipped with body surroundings ranging function, so that drivers can better understand changes in the surrounding environment and prevent traffic accidents. Traditional vehicle sensors include radar, laser, camera, etc., but they all have problems in detection accuracy and range [12]. With the development of vehicle networking technology, the application range of vehicle intelligent system is getting wider and wider. Binocular stereo vision sensor is a kind of localization and ranging algorithm based on parallax principle among vision sensors. The structure of the algorithm consists of five parts, which are image input, binocular localization, 3D correction, 3D matching and parallax calculation. The surface constraint-value domain scanning algorithm needs to be constructed before ranging to predict the distance in the real environment, and its threshold calculation method mainly refers to the threshold operation algorithm of the planar-space algorithm, as shown in equation (2) [13]

$$\begin{cases} x_{o1} = \dfrac{w}{2} \\ x_{o2} = \dfrac{3}{2}w + d \\ y_1 = f \end{cases} \tag{2}$$

In Eq. (2), x_{o1} and x_{o2} are the images acquired by the left and right sensors, respectively; f denotes the focal lengths of the left and right sensors; y_1 is the determinant of their focal lengths; d denotes the distance between the two sensors; and w denotes the width between the sensors. After inputting the parameters in the model, the detected target points must meet the projection requirements of the sensors. The projection function of the sensors is an imitation of the human eye visual system. The range of the binocular sensors is predicted by the surface constraint-value domain scanning algorithm, the left and right images are obtained from the predicted range, and the binocular localization function obtains the parameters and ensures that the left and right images remain at the same level. Then the SGBM stereo matching algorithm is used to obtain the parallax value of binocular ranging, the parallax value is the result of matching between the pixel points of the left and right images, which can be viewed as the difference between the different pixel points. The SGBM stereo matching algorithm for matching the pixel points includes four key steps, namely, image preprocessing, temporal cost computation, dynamic coordinate localization, and image post-processing, which are shown in Figure. 2 [14].

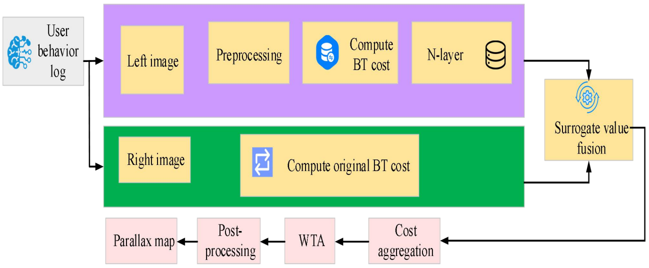


Figure 2. Flowchart of SGBM algorithm.

The image preprocessing of the SGBM algorithm is mainly to obtain the image gradient information, which lays the foundation for the pixel point matching of the left and right images. The pixel points of the preprocessed image are mapped by the mapping function to form a new image, and the new image pixel values are shown in equation (3).

$$P_{new} = \begin{cases} 0, P < -\alpha \\ P + \alpha, -\alpha \le P \le \alpha \\ 2\alpha, P \ge \alpha \end{cases} \tag{3}$$

In Eq. (3), P denotes the pixel value of the input image; P_{new} denotes the pixel value of the new image; α denotes a constant parameter. The calculation of the time cost requires the use of the absolute difference and the idea of domain summation of to calculate the local cost, the matching between the pixels of the left and right images also generates a cost, which is made up of the fusion of two parts. One part is the cost of the gradient of the image obtained by applying the sampling method to the input image preprocessing, and the other part is the cost of the initial block truncation coding obtained by using the sampling method on the initial image.

Since it will have high time cost to perform dynamic programming on 2D image and get the optimal solution, it is decomposed into simple 1D image problems to solve, and the dynamic programming method is applicable to all 1D image problems.SGBM stereo matching algorithm is a non-global matching algorithm, which can be constrained by using the global Markov energy equation for multiple path directions of the one-dimensional image, and it can effectively avoid the parallax information of one-dimensional image It can effectively avoid the cumulative effect of one-dimensional image parallax information, which passes the error information to the next path, and the screening of parallax of each pixel in the image is determined by the WTA. The last step is the image post-processing, this step is to check the uniqueness of the pixel points and pixel differences, and keep the synchronization of the detection of the left and right pixel points. During the image post-processing, if the curve smoothness is maintained when the gradient of the fitting function of the parallax value becomes small, the image edges and details can be preserved to the maximum extent. Therefore, a weighted least squares filter is used to achieve the smoothness of the curve when the gradient of the parallax value changes.

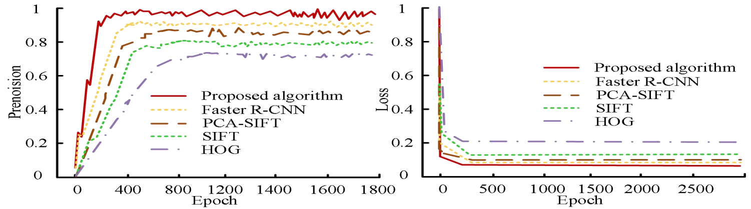
3. RESEARCH ON THE PRACTICAL APPLICATION OF BINOCULAR STEREO VISION RANGING METHOD

The improved Mask R-CNN algorithm shows some improvement in the five evaluation metrics of precision, recall, Dicc loss, mAP and F1 compared with the pre-improvement one, as shown in Table 1. In the table, A denotes FPN optimization, B denotes MobileNet-V3_MS optimization, C denotes PANet optimization; D denotes CA-SA optimization; and E denotes SGBM optimization. The FPN, MobileNet-V3_MS, PANet, CA-SA and SGBM optimizations of the algorithms improve the precision by 1.23%, 2.32%, 2.23%, 1.23%, and 1.23%, respectively; the recall improves by 2.34%, 0.12%, 1.23%, 2.13%, and 1.45%; and the Dicc loss improves by 2.23%, 1.13%, and 1.45%. 2.23%, 1.25%, 1.63, 1.23% and 0.12% respectively; mAP improved by 2.34%, 1.23%, 2.27%, 1.34% and 0.35% respectively; and FI improved by 2.24%, 0.54%, 1.25% and 0.34% respectively. Therefore the proposed method has better performance in detection performance after improvement compared to before improvement.

Table 1. Comparison of algorithm performance before and after optimization.

Model	P	Recall	Dicc losses	mAP	F1
Mask R-CNN+A	88.46	88.45	87.34	12.32	87.34
Mask R-CNN+A+B	90.01	91.56	86.12	13.67	86.12
Mask R-CNN+A+B+C	91.45	92.56	86.56	14.67	86.56
Mask R-CNN+A+B+C+D+E	91.65	94.35	89.50	16.5	89.50

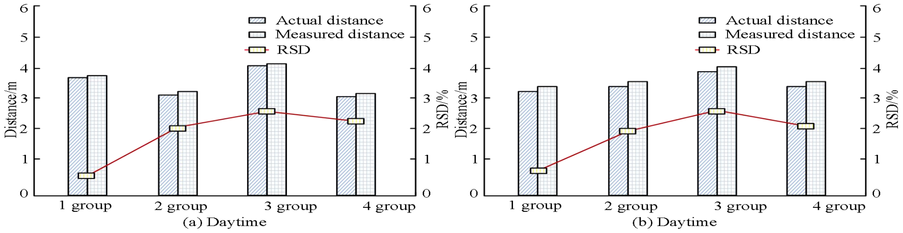
To further validate the performance of the proposed algorithm, it is understood that the performance of the algorithm is compared with other image processing algorithms of Faster R-CNN, PCA-SIFT, SIFT, and HOG in real ranging scenarios. From Figure. 3(a), it can be seen that the accuracy of the proposed algorithm tends to 97% after about 100 iterations; the accuracy of Faster R-CNN tends to 92% after about 200 iterations; the accuracy of PCA-SIFT algorithm tends to 90% after about 230 iterations; the accuracy of SIFT tends to 82% after about 260 iterations; the accuracy of HOG algorithm tends to 82% after about 380 iterations. 380 iterations or so, the accuracy rate tends to 80%. Compared to Faster R-CNN, PCA-SIFT, SIFT and HOG algorithms, the proposed algorithm outperforms the best performance in terms of detection accuracy, with accuracy improvements of 2.43%, 3.21, 2.12, 2.14% and 3.23%, respectively. In Figure. 3(b), it can be seen that with the increase in the number of iterations, both the proposed algorithm and the other image processing algorithms converge gradually, with the proposed algorithm having the best convergence. The superiority of the algorithm is verified.



(a) Accuracy curve (b) Loss curve

Figure 3. Curve of accuracy of different methods with the number of iterations.

To verify the actual detection effect of the proposed SGBM-based binocular stereo vision ranging method in real scenes, the two real scenes are daytime and nighttime. The proposed algorithm is applied to four different datasets, and the four datasets correspond to four different experimental groups, and the error between the actual measured distance and the true value is observed, and the results are shown in Figure. 4. In Figure. 4(a), it can be seen that the distance measured by the proposed algorithm and the difference between the true and actual measured values in all the 4 data sets are not much different and the errors are less than 5%. The error between the real and actual distances measured during the daytime is less than that measured at night, where the average relative error during the daytime between the four experimental groups is 3.54%, which is 0.35% less than the average relative error at night, which is due to the large image noise at night, which affects the sensor performance. The results show that the proposed algorithm has a good application in real application scenarios.



(a) Daytime (b) Daytime

Figure 4. Comparison of ranging errors in different scenarios.

4. CONCLUSION

In order to increase the detection and localization function of the vehicle network system, to improve vehicle intelligence and to reduce the incidence of road traffic safety accidents. The study proposes an optimized Mask R-CNN algorithm. The algorithm is optimized for FPN, MobileNet-V3_MS, PANet, CA-SA and SGBM the algorithm's precision, recall, Dicc loss mAP and F1 evaluation metrics are improved. The algorithm is optimized for FPN to improve its spatial and strong semantic information, and each evaluation metric is improved by 2.13% on average; the optimization of MobileNet-V3_MS optimizes the number of convolutional kernels and channels to achieve the algorithm lighter, and the evaluation metrics are improved by 2.03% on average; the optimization of PANet results in the shortening of the feature top-down transfer path, and the evaluation metrics are improved by 1.82%; the optimization of CA-SA enhances the attention share of image focus features, with an average improvement of 2.26% in each evaluation index; the optimization of SGBM module reduces the cost of algorithm computation time, with an average improvement of 2.43% in each evaluation index. The proposed algorithm achieves higher accuracy with fewer iterations than other Faster R-CNN, PCA-SIFT, SIFT, and HOG image processing algorithms, and after about 100 iterations, the accuracy tends to 97%, which is about 230 iterations fewer than the other algorithms on average, and the accuracy is improved by 2.34% on average. The proposed algorithm has a better application for binocular ranging of targets in vehicle networking system, which can provide some reference for the research of vehicle intelligent system. However, the enhancement of the image pixel resolution is still relatively small, and the algorithm rate can be further improved subsequently while ensuring the detection accuracy.

REFERENCE

[1] Zou Q, Sun Q, Chen L, Nie B, Li Q. A comparative analysis of LiDAR SLAM-based indoor navigation for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 6907-6921.

[2] Yasuda Y D V, Martins L E G, Cappabianco F A M. Autonomous visual navigation for mobile robots: a systematic literature review. ACM Computing Surveys (CSUR), 2020, 53(1): 1-34.

[3] Zhou X, Xu X, Liang W, Yan Z. Deep-learning-enhanced multitarget detection for end-edge-cloud surveillance in smart IoT . IEEE Internet of Things Journal, 2021, 8(16): 12588-12596.

[4] Dai Y, Wu Y, Zhou F, Barnard K. Attentional local contrast networks for infrared small target detection. ieee transactions on geoscience and remote sensing, 2021, 59(11): 9813-9824.

[5] Preethi P, Mamatha H R. Region-based convolutional neural network for segmenting text in epigraphical images. artificial intelligence and applications, 2023, 1(2): 119-127.

[6] Yu B, Bo M. Research on Small Target Ship Detection in Infrared Image based on YOLOV3. International Core Journal of Engineering, 2021, 7(9): 537-545.

[7] Lu X F, Bai X F, Li S X, Hei X. Infrared small target detection based on the weighted double local contrast measure utilizing a novel window. IEEE Geoscience and Remote Sensing Letters, 2022, 19(4): 1-5.

[8] Du J, Lu H, Hu M, Zhang L, Shen X. CNN-based infrared dim small target detection algorithm using target-oriented shallow -deep features and effective small anchor. iet image processing, 2021, 15(1): 1-15.

[9] Bi X, Hu J, Xiao B, Li W, Gao X. Iemask r-cnn: information-enhanced mask r-cnn. IEEE Transactions on Big Data, 2022, 9(2): 688-700.

[10] Sahin M E, Ulutas H, Yuce E, Erkoc, M. F. Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images. Neural Computing and Applications, 2023, 35(18): 13597-13611.

[11] Yu C, Hu Z, Li R, Xia X, Zhao Y, Fan X, Bai Y. Segmentation and density statistics of mariculture cages from remote sensing images using mask R-CNN. Information Processing in Agriculture, 2022, 9(3): 417-430.

[12] Droby A, Kurar Barakat B, Alaasam R, Madi B, Rabaev I, El-Sana J. Text line extraction in historical documents using Mask R-CNN. Signals, 2022, 3(3). 535-549.

[13] Hu G, Wang T, Wan M, Bao W, Zeng, W. UAV remote sensing monitoring of pine forest diseases based on improved Mask R-CNN. International Journal of Remote Sensing, 2022, 43(4): 1274-1305.

[14] Amo-Boateng M, Sey N E N, Amproche A A, Domfeh M K. Instance segmentation scheme for roofs in rural areas based on Mask R-CNN. The Egyptian Journal of Remote Sensing and Space Science, 2022, 25(2): 569-577.