Left and Right Consistent Stereo Image Detection and Classification Based on Deep Learning

Lijuan Tang¹; Qing Wang¹ and Guannan Chen¹,*

¹ Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Provincial Key Laboratory of Photonics Technology, Fujian Normal University, Fuzhou, China, Fuzhou, Fujian 350117, P. R. China
*email: edado@fjnu.edu.cn

Abstract. With the development of stereo camera technology, the increasing of the data volume of binocular image makes the target detection and classification of binocular image become one of the research directions in the field of computer vision. Compared with monocular image, binocular image contains the depth information of the image. For using the depth information of the image to improve the target detection and classification accuracy of image, this text puts forward a method of detecting and classifying left and right consistent objects in stereo images based on deep learning, which is based on the Mask R-CNN network and adds the binocular matching algorithm to improve the accuracy of image target detection and classification. The experiment compares the evaluation indexes of monocular and binocular.

1. Introduction
Stereo vision is an expression of the intuitive feeling of the distance between two eyes. In general, this feeling is three-dimensional. It vividly reflects the depth, unevenness and distance of objects in human binocular vision. Stereo vision technology is mainly divided into binocular vision technology and multi-eye vision technology[1]. Binocular technology vision is currently used more in stereo vision. The binocular stereo vision technology uses parallax theory to recover the depth information and three-dimensional coordinates of pixels. By acquiring two images acquired at the same time, the three-dimensional scene information is restored, the real three-dimensional world is restored, and the location information of the target is provided for navigation[2]. It is a kind of passive vision measurement technology. Binocular stereo vision technology is a research hotspot in the field of computer vision. The human vision system is simulated by a binocular camera to perceive the world. In public safety, robot navigation, object recognition, map generation, aerial exploration, seabed detection, 3D reconstruction and It has a wide range of applications in virtual reality.

As a branch of computer vision, binocular stereo vision occupies an important position in the academic field[3]. The binocular vision system mainly includes five basic modules: digital image pair acquisition, binocular stereo calibration, digital image pair correction, stereo matching, and three-dimensional structure restoration. The stereo matching module has always been the core of the system research. Stereo matching in binocular vision is actually establishing matching points between the two images, and then calculating the depth value in the scene based on the camera parameters and the corresponding matching points[4].

Stereo matching is one of the most important tasks in stereo vision algorithms. According to different matching primitives, stereo matching algorithms can be divided into feature-based matching algorithms, region-based matching algorithms and phase-based matching algorithms[5]. According to different matching strategies, stereo matching algorithms can be divided into local matching algorithms, semi-global matching algorithms and global matching algorithms. The local matching
algorithm mainly uses the local optimization method for parallax value estimation. The local stereo matching algorithm includes SAD, SSD and other algorithms. Like the global stereo matching algorithm, the parallax estimation is also performed by the energy minimization method. However, in the energy function, there are only data items and no smoothing items. Both the linear growth matching algorithm and the adaptive window algorithm are local matching algorithms. The semi-global stereo matching algorithm (SGM) is based on a pixel-by-pixel matching method, which uses mutual information to evaluate the matching cost and approximates a global two-dimensional smoothing constraint by combining many one-dimensional constraints[6]. The global stereo matching algorithm mainly uses a global optimization theory method to estimate disparity, establish a global energy function, and obtain an optimal disparity value by minimizing the global energy function. The main algorithms for global matching include graph cuts, belief propagation, and dynamic programming[7].

From the result of Mask R-CNN running, we can see that Mask R-CNN cannot detect the small and occluded objects[8]. For solving this problem, this text put forward a method of left and right consistent target detection of stereo image based on deep learning. This method adds binocular stereo matching algorithm based on Mask R-CNN network, whose aim is to solve the problem that the Mask R-CNN module could not detect small objects and blocked objects. This experiment uses the kitti binocular dataset to train and then selects out target detection and classification results between monocular and binocular images to compare. In this experiment, map (average accuracy) and AP (target detection and classification accuracy) of monocular and binocular images are compared respectively by setting evaluation indexes, as well. Another section of your paper

2. Methods

2.1. Semi-Global Stereo Matching

Semi-global stereo matching is a matching algorithm between local and global matching. It better neutralizes the advantages and disadvantages of local and global matching, and has a good balance in accuracy and efficiency. In fact, semi global matching still adopts the idea of optimizing the energy function in the global matching algorithm, that is, finding the optimal parallax of each pixel to minimize the global energy function of the whole image[9]. In fact, semi global matching still adopts the idea of optimizing. The energy function adopted by SGM is:

\[ E(D) = \sum_P (C(P, D_p)) + \sum_{q \in N_P} P_1 T \| D_p - D_q \| = 1) + \sum_{q \in N_P} P_2 T \| D_p - D_q \| > 1 \]  

Where D is the disparity map, \( E(D) \) is the energy function corresponding to the disparity map. \( p, q \) represents a pixel in the image. \( N_P \) refers to the adjacent pixels of pixel p. \( C(P, D_p) \) refers to the cost of the pixel when the disparity of the current pixel is \( D_p \). \( P_1 \) is a penalty coefficient, which applies to those pixels where the disparity value of the adjacent pixels of pixel p is different from the disparity value of p by 1. \( P_2 \) is a penalty coefficient, which is applicable to those pixels in which the disparity value of the adjacent pixels of p is different from the disparity value of p by more than 1. Where \( P_2 > P_1 \), A smaller penalty term allows the algorithm to adapt to situations where the parallax changes are small, such as inclined planes or continuous surfaces[10]. A larger penalty term allows the algorithm to correctly handle parallax discontinuities. Because the possibility of discontinuity at the edge of the grayscale of the image is greater[11], in order to protect the discontinuity in the real scene, is often dynamically adjusted based on the grayscale difference of adjacent pixels, as shown in Equation 2:

\[ P_2 = \frac{P_2'}{I_{bp} - I_{bq}} \quad P_2 > P_1 \]
\( p_2 \) is the initial value of \( p_2 \).

\( p_2 \) the optimization of the energy function in Equation 1 is an NP problem. To solve it efficiently, SGM proposes a path cost aggregation. That is, the matching costs under all disparities of a pixel are aggregated one-dimensionally on all paths around the pixel to obtain the path cost value under the path, and then all path cost values are added to obtain the matching cost value after the pixel is aggregated.

2.2. Semi-Global Stereo Matching Based on Mask R-CNN

Mask R-CNN is based on Faster R-CNN, Faster R-CNN is a typical two-stage target detection method. First, RPN candidate regions are generated, and then the candidate regions undergo target detection (including target classification and coordinate regression) through Roi Pooling. Classification and regression share the previous network. Mask R-CNN is also two-stage, and the RPN part is the same as Faster R-CNN[12]. Then, Mask R-CNN adds a third branch based on Faster R-CNN, and outputs a Mask for each Roi. The structure of Mask R-CNN mainly includes a shared convolution layer, a candidate region generation network (RPN), a network that classifies candidate regions and generates a mask (three branches).

Based on the Mask R-CNN network, the solution adds semi global stereo matching module. We have known that the target detection and classification effect of Mask R-CNN is very good. How to further improve the detection and classification accuracy of Mask R-CNN is a difficult problem. Through looking up a lot of literature, the text proposes to add binocular matching to improve the detection and classification of Mask R-CNN. Because Mask R-CNN detects and classifies single-purpose data sets, it’s hard to use depth information of image. Therefore, the text proposes binocular stereo matching algorithm added. Semi global stereo matching is good at neutralizing the advantages and disadvantages of local matching and global matching, so this paper adds semi global stereo matching algorithm. The network framework of this paper is shown in figure 1. Compared with figure 1, it can be seen intuitively that this experiment added a semi-global stereo matching module to the network framework of Mask R-CNN.

![Figure 1. The network framework of Semi-Global Stereo Matching Based on Mask R-CNN](image-url)
3. Result
In this experiment, we perform object detection, classification, and segmentation experiments on the binocular dataset. When performing target detection and classification experiments on monocular images, we only need to use the left or right image in the kitti binocular dataset for training. When performing object detection and classification experiments on binocular images, we need to input left and right images for training at the same time. In order to evaluate the performance advantages of adding the semi-global stereo matching algorithm, we selected several comparatively obvious result maps for comparison, and compared the average accuracy and target detection and classification accuracy of monocular images and binocular images. Some results compared with monocular and binocular were shown on Figure 2.

![Figure 2](image-url)

**Figure 2.** The results were compared with monocular and binocular
From the monocular and binocular result maps, we can clearly contrast that binocular can detect smaller objects and covered objects. For example, (a) in the figure, binocular detects more handbags and vehicles behind the vehicle than monocular. In (b), binocular eyes detect more planes and cars than monocular eyes. In the figure (c), a bench is detected for the binocular target detection, but a bench is not detected for the monocular detection. However, we can also know from the prediction probability of the monocular and binocular result graphs that the monocular prediction probability will be higher than that of the binocular.

**Table 1. Comparison of monocular and binocular average detection accuracy**

| model       | method | map |
|-------------|--------|-----|
| Mask r-cnn  | Monocular | 0.87 |
| Mask r-cnn+SGM | Binocular | 0.93 |

Table 1 shows the comparison of monocular and binocular detection methods. By observing the detection accuracy in Table 1, we can find that the average detection accuracy of binocular is 0.4 points higher than the monocular.

**Table 2. Comparison of Monocular and Binocular Target Detection and Classification Accuracy**

| model       | method | AP | AP_50 | AP_75 | AP_s | AP_m | AP_l |
|-------------|--------|----|-------|-------|------|------|------|
| Mask r-cnn  | Monocular | 32.9 | 53.6 | 32.1 | 12.5 | 35.2 | 48.6 |
| Mask r-cnn+SGM | Binocular | 34.8 | 56.9 | 36.5 | 14.6 | 38.2 | 50.9 |

Table 2 shows the comparison of monocular and binocular target detection and classification accuracy. By observing the target detection and classification evaluation indicators in Table 2, we know that adding the binocular stereo matching algorithm improves the AP by about 2 points, increases AP_50 (estimated the average accuracy by 0.5 as the IoU threshold), and increases AP_75 (by 0.75 is the IoU critical value to estimate the average accuracy) by 4.4 points. In short, the binocular AP (average value of the IoU threshold), AP_50, AP_75, and AP_s, AP_m, and AP_l are higher than those of the monocular.

In this experiment, we detect by binocular instead of monocular increasing input characteristics to gain more comprehensive image information. Because of more effective expression of foreground and background information that is segments by binocular image, it can provide supplementary information for color and texture, and can solve the accuracy and integrity that cannot be solved by monocular image detection. We add the binocular stereo matching algorithm to better detect small objects and occluded objects, which improves the detection accuracy at the same time.

**4. Conclusion**

In this paper, we propose adding binocular stereo matching algorithm to improve the accuracy of image detection and classification in Mask R-CNN network. First, we put the semi global stereo matching module behind the resnet-101 module. Nest, training for the model. Finally, Comparing the results of monocular and binocular training. We can find that adding semi global stereo matching algorithm can get higher evaluation index.

5. References

[1] Geiger A, Roser M and Urtasun R. 2010. *In Tenth Asian Conference on Computer Vision*. 
  Springer, Berlin, Heidelberg, pp 25-38.

[2] Jie L, Zhang J, Yu D and He S. 2018. *Acta Oprica Sinica*, 38, 0115004.
[3] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A. C. and Fei-Fei, L. 2015 *International Journal of Computer Vision* pp 211-252.

[4] Hirschmüller H, Innocent P. R and Garibaldi J. 2002 *International Journal of Computer Vision* pp 229-246.

[5] Žbontar J and LeCun Y. 2006 *J. Mach. Learn. Res.* pp 2287–2318.

[6] Ren S, He K, Girshick R and Sun J 2017 IEEE Transactions on Pattern Analysis and Machine Intelligence pp 1137-1149.

[7] Kaiming H, Georgia G, Piotr D, and Ross G 2018 IEEE Transactions on Pattern Analysis and Machine Intelligence p 1

[8] Einecke N and Eggert J. 2015 Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV) pp585-592.

[9] Shelhamer E, Long J and Darrell T. 2017 IEEE Transactions on Pattern Analysis and Machine Intelligence pp640-651.

[10] Hariharan B, Arbeláez P, Girshick R and Malik J 2014 *Proceedings of the European Conference on Computer Vision* pp 297-312

[11] Ioffe S and Szegedy C. 2015. Proceedings of the Proceedings of the 32nd International Conference on International Conference on Machine Learning pp 448–456.

[12] Bell S, Lawrence Zitnick, C, Bala K and Girshick R 2016 *The IEEE Conference on Computer Vision and Pattern Recognition* pp2874-2883.