Three-dimensional Spatial Localization Based on Binocular Vision

Lan Zang

Hainan University <https://orcid.org/0000-0002-2307-6644>

Kun Zhang (✉ kunzhang@hntou.edu.cn)

Hainan University <https://orcid.org/0000-0001-9195-8000>

Chuan Tian

Sanya Institute of Deep-sea Science and Engineering Chinese Academy of Sciences

Chong Shen

State Key Laboratory of Marine Resource Utilization in South China Sea

Bhatti Uzair Aslam

Nanjing Normal University School of Geography

Ju Huang

Hainan University

Research

Keywords: binocular vision, object detection, Faster R-CNN, three-dimensional space positioning

Posted Date: November 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1067324/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Three-dimensional Spatial Localization Based on Binocular Vision

Lan Zang^{1,3}, Kun Zhang^{1,2,*}(Member, IEEE), Chuan Tian^{4,*}, Chong Shen^{1,3,*}(Member, IEEE), Bhatti Uzair Aslam⁵, Ju Huang^{1,3}

¹ State Key Laboratory of Marine Resources Utilization in South China Sea, Hainan University, Haikou, Hainan, 570228, China

² Education Center of MTA, Hainan Tropical Ocean University, Sanya, Hainan, 572022, China

³ School of Information and Communication Engineering, Hainan University, Haikou, Hainan, 570228, China

⁴ Institute of Deep-sea Science and Engineering, Chinese Academy of Science, Sanya, Hainan, 572000, China

⁵ School of Geography, Nanjing Normal University, Nanjing, Jiangsu, 210023, China

* Correspondence: Kun Zhang (e-mail: kunzhang@hntou.edu.cn), Chuan Tian(e-mail: tianc@idsse.ac.cn), Chong Shen (e-mail: chongshen@hainanu.edu.cn).

This work was supported by the National Natural Science Foundation of China (No. 61861015); 2021 Hainan Province's Major Science and Technology Plan Project "Multimodal Medical Big Data Research and Application Demonstration of Common Key Technologies for Diagnosis and Treatment of Special Diseases" under Grant (No. 202149); 2021 Key Research and Invention Projects of Hainan Province (No. ZDYF2021GXJS032); the Open Project of the State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University under Grant (No. MRUKF2021032).

Abstract: With the rapid innovation of science and technology, researchers are no longer satisfied with the simple reconstruction and composition of binocular vision. Therefore, in order to solve the problems of low accuracy and unstable system performance, this paper proposes a three-dimensional space recognition and location algorithm based on binocular stereo vision and deep learning algorithm. Firstly, Zhang's calibration method is used to set the calibration error at 0.10 pixels, and sad algorithm is selected to reduce the search range of matching points and reduce the data burden for subsequent experiments. Then, the three-dimensional spatial data calculated by binocular parallax is input into Faster R-CNN model for data training, and the target feature is extracted and classified. Finally, the object and its position coordinate information can be detected in real time. Experimental data analysis shows that when the calibration error is the best, and the number of data training is enough, the algorithm can effectively improve the quality of target detection, positioning accuracy and target recognition rate is improved by about 3% - 5%, and can achieve faster fps.

Keywords: binocular vision; object detection; Faster R-CNN; three-dimensional space positioning

I. Introduction

Global positioning system, Beidou system, Galileo system and other positioning systems are playing an important role in more and more fields by virtue of their high positioning accuracy and accurate time service information function. However, due to the complexity of the field environment or indoor environment, GPS and

other positioning systems can not work effectively [1-2]. Indoor location research novel coronavirus pneumonia has been developed. In 2020, the new crown pneumonia epidemic was raging. The indoor mobile robots with various functions helped doctors and nurses to complete the work of transmitting and disinfecting medical products, and were welcomed by all sectors of society, especially the medical profession. The indoor positioning technology based on vision slam relies on stereo vision as the main means of positioning and navigation. It does not need radio signal transmission, nor does it need infrared sensors, lidar and other ranging sensors. It is a non-contact ranging scheme, which not only solves the problems of signal interference and non line of sight error, but also reduces the application cost [3-4].

Three-dimensional space positioning based on binocular vision is a positioning method developed in recent years. It uses binocular cameras to obtain images, and then uses computer vision algorithm for image processing to calculate the three-dimensional information of the scene, and obtains the exact location information of the target object through recognition, so as to achieve the effect of three-dimensional space positioning of the object [5-6]. Visual slam generally goes through image acquisition, camera calibration, feature extraction, feature matching and other links [7]. Researchers also use various improvement methods to further improve the high accuracy and robustness of the visual positioning system. Camera calibration is to understand the transformation relationship between the object from the real world to the computer image plane.

Zhang Zhengyou plane calibration method is the most commonly used one [8]. In order to obtain accurate depth information, the calibration accuracy is improved by improving the checkerboard template, OpenCV coupling Zhang Zhengyou calibration method and other improved methods [9-12]. After the completion of target calibration, feature points are generally extracted. Common point feature-based algorithms include SIFT, SURF, ORB and so on. Researchers deeply study the extraction and matching of feature points and realize the accuracy and real-time performance of the algorithm [13-17]. In [18], a variety of vision schemes are proposed for different complex environments. In [19], binocular vision is applied to the design of autonomous mobile robot. The visual positioning technology is quite mature.

With the rapid development of artificial intelligence, deep learning algorithm has made breakthrough progress in the field of machine vision. The [20-23] summarizes the deep learning technology, and proposes that it is an inevitable trend to apply deep learning to visual field in the future. There are two categories of binocular vision algorithms based on deep learning: classification based and regression based. In order to meet the requirements of indoor high accuracy, classification based algorithms are usually selected [24-26]. In order to improve the system performance, [27] innovatively proposed Faster R-CNN algorithm, [28] used Fast R-CNN training data and surf algorithm for stereo matching. On this basis, [29] fused Faster R-CNN with multi-scale features, and [30] proposed an improved deep learning target detection framework Faster R-CNN model to improve the detection effect and positioning accuracy. Then, by comparing different data sets for training and testing, we can find a more accurate and time-consuming data model, and achieve effective indoor recognition and three-dimensional spatial positioning of objects [31-35].

Based on binocular stereo vision, this paper proposes a three-dimensional space positioning algorithm combining depth learning and binocular vision, which can output the coordinates of objects in real time and effectively improve the recognition accuracy of the target. The main contributions of this paper are as follows.

1) This paper proposes a block stereo matching algorithm with fast search speed, which helps to reduce the

search range of matching points, thus reducing the amount of computation, so as to achieve higher FPS.

2) A binocular target detection method based on deep learning algorithm is proposed. Compared with using binocular recognition alone, the positioning accuracy can be improved by 3% - 5%.

The rest of the article is as follows. The second part introduces the theory of binocular Stereo Vision and the algorithm of object detection based on depth learning, and summarizes the design of binocular vision and depth learning. The experiments and analysis are described in section III and the conclusions in section IV.

II. Method

2.1 Binocular stereo vision

In this section, we will briefly introduce the binocular camera model, camera calibration and stereo matching, which are needed in the follow-up knowledge.

2.1.1 Coordinate system

Before analyzing the camera model, we first have a brief understanding of the camera coordinate system, which lays the foundation for subsequent experiments and analysis, and is also the auxiliary measurement of three-dimensional reconstruction. Binocular coordinate system generally includes world coordinate system, camera coordinate system and image coordinate system. Among them, the world coordinate system is used as the reference system of the target object in binocular vision, and from this step, the target object is included in the operation. Camera coordinate system is the coordinate system to measure the camera's own angle, which is the only way to transform the world coordinate system to the image coordinate system. The object in the world coordinate system is transferred to the camera coordinate system through rigid body transformation, and then the position coordinates of the object in the image coordinate system are obtained by perspective projection. The image coordinate system is the representation of the object in the image [4].

Figure 1. Coordinate System Diagram

Due to the invariance of rigid body transformation, there will be no deformation when using rigid body transformation between world coordinate system and camera coordinate system, and the relationship between them is expressed by the following formula:

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} + \begin{bmatrix} T_X \\ T_Y \\ T_Z \end{bmatrix} \quad (1)$$

The homogeneous expression can be written as:

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = \begin{bmatrix} R & t \\ 0_3^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 & t \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 0 \\ 1 \end{bmatrix} \quad (2)$$

$$= \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix}$$

Among them, R is the unit orthogonal matrix, t is the translation vector, and R 、 T is actually rotation, translation and other operations.

The transformation from camera coordinate system to image coordinate system is a changing relationship from 3D to 2D. The relationship between x and y can be known by using similar triangles:

$$Z_C \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \quad (3)$$

$$= \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix}$$

Where, f is the focal length of the camera.

To sum up, the world coordinate system can be converted to the image coordinate system by using the matrix P :

$$P = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} r_{00} & r_{01} & r_{02} & T_X \\ r_{10} & r_{11} & r_{12} & T_Y \\ r_{20} & r_{21} & r_{22} & T_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

2.1.2 binocular camera model

In the vision slam system, the camera is usually used to obtain the information in the scene, and then the position,

geometry, height, color and other information can be used for corresponding processing. The binocular camera is composed of two monocular cameras, and the distance between the two cameras has been determined. The inspiration of binocular camera comes from human eyes. When human eyes observe things, they will have a three-dimensional sense of things because of their own depth perception ability. Therefore, the left and right cameras of binocular camera, like a pair of eyes of the same person, are in the same plane and the optical axes taken are parallel to each other [4]. Binocular vision can calculate the three-dimensional coordinates of each pixel in three-dimensional space through parallax, and when calculating the parallax between two images, it can directly measure the distance of the object in front, without judging what type of obstacles exist in front [5]. It is mainly based on two cameras fixed in different positions, through one or some features of the image of the object to calculate, respectively obtain the object's coordinate position on the two image planes. The model of binocular camera is shown in figure 2:

Figure 2. Binocular Camera Model

We can't calculate all the captured points, so we need to extract some characteristic points from them for auxiliary calculation. According to the above model, as long as we can determine the relative position of the camera, and then use the geometric relationship, we can calculate the coordinate position of the feature point in a certain stereo camera coordinate system, however, the first thing to be clear is that for the left and right cameras, the measuring point must correspond to the same, that is, the left and right cameras are in the same coordinate system [4-6].

In this space coordinate system, let the left camera be the main camera, then the object point in its world coordinate system and the camera coordinate system of the left camera will be transformed as shown in equation 5:

$$\begin{bmatrix} X_l \\ Y_l \\ Z_l \\ 1 \end{bmatrix} = \begin{bmatrix} R_l & \vec{t}_l \\ 0 & 1 \end{bmatrix}_{4 \times 4} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (5)$$

After the camera is calibrated, the position relationship between the right camera and the left camera can be calculated, as shown in equation 6:

$$\begin{bmatrix} X_r \\ Y_r \\ Z_r \\ 1 \end{bmatrix} = \begin{bmatrix} R & \vec{T} \\ 0 & 1 \end{bmatrix}_{4 \times 4} \begin{bmatrix} X_l \\ Y_l \\ Z_l \\ 1 \end{bmatrix} = \begin{bmatrix} R & \vec{T} \\ 0 & 1 \end{bmatrix}_{4 \times 4} \begin{bmatrix} R_l & \vec{t}_l \\ 0 & 1 \end{bmatrix}_{4 \times 4} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (6)$$

$$\text{Where, } \begin{bmatrix} R_r & \vec{t}_r \\ 0 & 1 \end{bmatrix}_{4 \times 4} = \begin{bmatrix} R & \vec{T} \\ 0 & 1 \end{bmatrix}_{4 \times 4} \begin{bmatrix} R_l & \vec{t}_l \\ 0 & 1 \end{bmatrix}_{4 \times 4}.$$

The rotation matrix and translation vector of the right camera relative to the measuring point can be obtained by the above formula derivation [15].

2.1.3 Camera calibration

Camera calibration is to understand the transformation relationship between the object from the real world to the computer image plane, and solve the internal and external parameters. In addition, the camera perspective projection will be distorted due to various deviations. Therefore, the calibration is also to solve the distortion coefficient for image correction.

Because of the distortion of the image taken by the camera, we should understand the distortion before calibration. Radial distortion and tangential distortion are usually considered in distortion. Radial distortion is the distortion of light in radial position. Generally, the radial distortion in the center of image plane is 0, and the distortion becomes more and more serious as it moves to the edge. Taylor series expansion can be used to describe its mathematical model [10]:

$$x_0 = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \quad (7)$$

$$y_0 = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \quad (8)$$

Where, (x_0, y_0) represents the original position of the distortion point, r represents the distance between the point and the center of the image plane, (x, y) is the new position after correction.

The tangential distortion is the distortion caused by the image plane and the lens are not completely parallel. The model can be described by the following formula and adding two additional parameters p_1 and p_2 [17]:

$$x_0 = x + [2p_1 y + p_2(r^2 + 2x^2)] \quad (9)$$

$$y_0 = y + [2p_2 x + p_1(r^2 + 2y^2)] \quad (10)$$

Zhang Zhengyou's calibration method is generally used in camera calibration, which was proposed by Professor Zhang Zhengyou in 1998 [8]. It is a single plane chessboard method based on coplanar spatial feature points. The calibration method uses the checkerboard calibration board shown in Figure 3, to print and paste on a plane, the image is collected by changing the angle and direction of the calibration plate, and the feature points in the image are extracted to solve the internal and external parameters of the camera in the ideal distortion free state, and the actual radial distortion coefficient is solved by using the least square method. The estimation accuracy is improved by the maximum likelihood method, and the internal parameters, external parameters and distortion coefficient of the camera are obtained [9-12]. The Zhang's calibration method used in this paper is convenient and accurate without the aid of special calibrators. The internal parameter matrix M_r and distortion matrix kp_r of the right camera obtained by calibration are as follows:

$$M_r = \begin{bmatrix} 488.6754 & -0.3216 & 299.2910 \\ 0 & 486.0078 & 244.7685 \\ 0 & 0 & 1 \end{bmatrix} \quad (11)$$

$$kp_r = [0.0630 \ 0.4040 \ -0.0032 \ 0.0062 \ -1.2823] \quad (12)$$

The internal parameter matrix M_l and distortion matrix kp_l of the left camera are as follows:

$$M_l = \begin{bmatrix} 489.3439 & -0.6205 & 346.4286 \\ 0 & 487.4673 & 249.7389 \\ 0 & 0 & 1 \end{bmatrix} \quad (13)$$

$$kp_l = [0.0757 \ 0.2405 \ -0.0052 \ 0.0081 \ -0.7321] \quad (14)$$

Then the rotation matrix R and translation relation vector T between the two cameras are obtained as follows:

$$R = \begin{bmatrix} 1.0000 & 0.0010 & -0.0003 \\ -0.0010 & 1.0000 & 0.0044 \\ 0.0003 & -0.0044 & 1.0000 \end{bmatrix} \quad (15)$$

$$T = [60.1051 \quad 0.0761 \quad 0.5573] \quad (16)$$

Figure 3. Chessboard Calibration Board

2.1.4 Stereo matching

Stereo matching is a key step in binocular vision. It is to match the matching points of left and right images and the points to be matched in the same frame after stereo correction, and then get the two-point disparity. Then, the depth information of the image is calculated according to the parallax relationship. Therefore, it determines the accuracy of the whole algorithm. Common stereo matching algorithms include gray based matching algorithm, feature based matching algorithm and relation based matching algorithm. In this paper, we choose the template matching algorithm based on gray level, that is, the two camera frames are divided into many small squares. Using the known template image, we move the small squares to match with the small squares in another image, compare the position relationship of the two pixels, and calculate the actual depth of the object. Moreover, most of the matching algorithms should comply with four criteria: epipolar constraint, uniqueness constraint, parallax continuity constraint and sequence consistency constraint [13-17]

In block stereo matching BM algorithm, the following similarity measurement function can be used for calculation:

(1) Sum of Absolute Differences (SAD) :

$$D(i, j) = \sum_{s=1}^M \sum_{t=1}^N |S(i+s-1, j+t-1) - T(s, t)| \quad (17)$$

(2) Sum of Squared Differences (SSD) :

$$D(i, j) = \sum_{s=1}^M \sum_{t=1}^N [S(i+s-1, j+t-1) - T(s, t)]^2 \quad (18)$$

(3) Normalized Cross Correlation (NCC) :

$$R(i, j) = \frac{\sum_{s=1}^M \sum_{t=1}^N [S^{i,j}(s, t) - E(S^{i,j})] \cdot [T(s, t) - E(T)]}{\sqrt{\sum_{s=1}^M \sum_{t=1}^N [S^{i,j}(s, t) - E(S^{i,j})]^2 \cdot \sum_{s=1}^M \sum_{t=1}^N [T(s, t) - E(T)]^2}} \quad (19)$$

In this paper, SAD algorithm is used to sum the absolute value of the corresponding value difference of each pixel, and then compare the similarity, which is a local matching algorithm[17]. The basic process is as follows: define a window, take the matching point of the left view as the center point, count the sum of the gray values of the window in the left view, use the same steps to obtain the gray sum of the right view, then subtract the right view from the gray sum of the left view, and repeat the calculation step by step. Finally, search the area with the smallest difference within this range, which is the best matching block. SAD algorithm is fast and simple, easy to parallel, and has good real-time performance. It can effectively search for a wide range of motion of different blocks, and also helps to reduce the search range of matching points. The disparity between the left and right images can be obtained by sad algorithm, and the depth map can be obtained by further processing based on the inverse relationship between depth and disparity.

2.2 Object detection based on deep learning

In this section, we mainly introduce the object detection algorithm based on deep learning, and briefly introduce the key modules of the faster R-CNN model.

As the core problem of computer vision, object detection plays an important role in three-dimensional positioning. It is necessary to apply deep learning to the field of vision [20-21]. Object detection algorithms based on deep learning can be divided into two categories: one is based on classification, such as R-CNN series model; the other is based on regression, such as YOLO model, SSD model and so on [22-23]. The reason why this paper chooses the algorithm based on classification is that this kind of target detection algorithm has higher accuracy and is more suitable for indoor places with higher accuracy requirements.

The main idea of the Faster R-CNN algorithm is to generate several candidate frame regions by using the preset anchor frame style, and then carry out subsequent

operations such as decision. Faster R-CNN is mainly divided into four parts : 1) conv layers, as the extraction of image features, the image is input into convolution network, and the corresponding feature mapping is generated; 2) Region Proposal Network (RPN), as a candidate region for recommendation, replaces the previous search selective to generate proposals, so that subsequent classification, regression and other operations share convolution features. Because RPN and fast R-CNN share a CNN [23], the input here can be regarded as convolution layer feature maps; 3) ROI pooling, the feature map with the same size is generated from the input of different sizes through the ROI pooling layer; 4) Classification and accurate location are the ultimate goal of the whole system. Candidate frames of target objects are generated and the exact location information of candidate frames in the image is output [22]. The overall flow of Faster R-CNN target detection is shown in figure 4:

Figure 4. Faster R-CNN Detection Framework

Where, the core components of Faster R-CNN include RPN candidate box extraction module and Fast R-CNN detection module.

2.2.1 Anchor

Anchor is the core concept of RPN, so what is anchor? The essence of anchor is the mapping point of the center point of the current sliding window on the feature map on the original image, which emphasizes that anchor is the point in the original region. Anchor mechanism is the reverse thinking of spatial pyramid pooling (SPP). SPP can transform feature maps of any size into feature vectors of fixed size. Therefore, anchor reverses the outputs of the same size to obtain inputs of different sizes.

In target detection, anchor represents fixed reference frame, anchor-based can be divided into single-stage and two-stage detectors. Faster R-CNN lays the dominant position of anchor-based two-stage detector. In the first stage, the input image and the suggested region of the object are browsed, and in the second stage, the output region is classified. Anchors is a set of reference windows

containing $\{128 * 128, 256 * 256, 512 * 512\} * \{1:1, 1:2, 2:1\}$, whose dimensions and aspect ratio are fixed. In RPN network, there are nine possibilities corresponding to each sliding window position in the original image area, which are expressed as nine templates with different scales and aspect ratios. Thus, the detection problem is transformed into whether the target object exists in the reference frame and the distance from the reference frame. In the RPN network, a set of fixed reference frames with different scales and positions are set, and the possible location regions of objects are determined by using anchors mechanism. Multiple anchor boxes are obtained by using different scales and aspect ratios in the regions, and then refined by CNN. We can see the appearance of anchor many times in RPN network. The head is the set reference frame, the middle part carries out the task calculation of classification and border regression, and the tail part is the summary of the calculation results of the middle part, which carries out the preliminary screening and preliminary offset of anchor. Because the detection algorithm is two stage type, there will be two tasks of classification and border regression in the middle part. Anchor solves the multi-scale problem and changes the predicted value from the absolute coordinate to the predicted offset, which makes the training easier to converge and improves the accuracy of the detector.

2.2.2 RPN Network

RPN discards selective search and directly uses convolutional neural network CNN to generate detection frames of candidate regions, which is the core of the whole network. In fact, RPN network has two lines: softmax classifiers to obtain positive and negative, and bounding box region offset to calculate relative anchors to obtain accurate proposal. It is a full convolution network, which can judge whether it is a target when predicting the target area box of each position in the input image [25].

RPN is composed of a $3 * 3$ convolution layer and two $1 * 1$ convolution networks. Firstly, the convolution operation of the basic network is used to get the common feature map and input it into RPN network. In the RPN stage, they will go through a $3 * 3$ convolution layer to get $256 * 16 * 16$ feature graph, which makes the anchor points on the feature graph generate anchor box with inconsistent size and aspect ratio, and then input the generated anchor box into two

parallel $1 * 1$ convolution networks. With the help of preset anchors, one performs the classification task to analyze whether there is a target in the candidate box, and the other completes the acquisition of frame position information to achieve the purpose of preliminary positioning [29].

RPN first appeared in the network structure of Faster R-CNN. Its essence is "classless object detector based on sliding window". Its main task is to generate proposals and use them together with the feature map of the last layer. In fact, RPN is used to detect objects and distinguish whether the objects in the image are objects or backgrounds. In RPN, anchors are used to uniformly place a fixed size reference bounding box in the pixel space of the original image to transform the problem into the thinking of anchor. The proposed RPN network can integrate the whole target detection process into the neural network in a real sense, get rid of the selective search algorithm, and greatly improve the recognition speed.

2.2.3 Faster R-CNN detection module

Fast R-CNN is an improvement of R-CNN, which makes fine-tuning of some convolution layers in the network, so that the detection effect is better. The workflow of Fast R-CNN network is as follows: firstly, the segmentation method is used to select 2K candidate boxes in the image selective search as the input, then several convolution layers and maximum pooling layer are used to process the image, that is, after obtaining the feature mapping, the image features are extracted by convolution neural network and the convolution feature map is generated. According to the corresponding region selected by the previous ROI box, the part of the candidate box is sampled before the last convolution operation, and the ROI pooling layer is used to unify the proportion of feature vectors in the feature map. Finally, these eigenvectors are input into a series of fully connected layers, which are finally divided into two output layers of the same level: one is used to predict the category, and the other is on the linear regressor to adjust the position of the bounding box. The former outputs soft-max probability estimates of 21 categories, while the latter outputs 4 real values of 4 coordinates of each region proposal, and the 4 values of each group represent the correction of the detection frame position of each category [28].

Fast R-CNN skillfully integrates the classification and

regression tasks into one model, which avoids the problem of repeated feature extraction and calculation, and only needs one feature extraction for the whole image; at the end of the network, a parallel full connection layer is adopted, so that the classification results and window regression results can be output at the same time, reducing the extra storage space of features. Compared with R-CNN, fast R-CNN has a significant improvement in detection speed and accuracy.

TABLE I. COMPARISON of RCNN, FAST RCNN and FASTER RCNN

	Accuracy	Test Time(seconds)	Speedup
R-CNN	66%	47	1X
Fast R-CNN	70%	3	25X
Faster R-CNN	73.2%-85.6%	0.2	250X

R-CNN series algorithms are target detection algorithms based on classified convolutional neural network structure, which is a new milestone in the application of CNN method to target detection. R-CNN was proposed in 2014. However, as there are as many as 2000 candidate frames, the amount of computation is too heavy, so it takes 47s for R-CNN to detect an image. Fast R-CNN is mainly to accelerate the R-CNN processing, from the original 47s to 3s, but also to ensure the accuracy and speed of multiple detection methods. Faster R-CNN, as the best detection algorithm of RCNN series, can be regarded as a system of "RPN + Fast R-CNN". Not only the single layer SPP net is added as ROI Pooling layer on the basis of R-CNN, and the regional recommendation network is proposed based on Fast R-CNN. In technology, the RPN network and Fast R-CNN network are combined to realize the sharing of the feature weights of the two, so that the target detection can be realized by the neural network to complete the real end-to-end training [42]. RPN network has high efficiency in generating ROI, and can run every image at 10 ms speed, and the speed of target detection can reach 17fps in simple network and 5fps in complex network. Faster R-CNN algorithm creatively uses convolutional network self suggestion boxes, which reduces the number of suggestion boxes from the original 2000 to 300, and improves the quality of suggestion boxes. And the feature extraction, region recommendation, boundary regression and classifier are all done by the deep neural network, and run on GPU, not only the network performance is more stable, but also greatly improve the

speed of target detection. Anchor solves the multi-scale problem and changes the predicted value from the absolute coordinate to the predicted offset, which makes the training easier to converge and improves the accuracy of the detector.

2.3 Experimental design

Binocular vision mainly uses the "parallax" of the left and right images, and restores the three-dimensional information of the environment according to the geometric principle. In the three-dimensional reconstruction, depth learning algorithm is introduced to complement each other and improve the efficiency. The specific steps are to use binocular to calculate three-dimensional spatial data, and complete operations such as distortion removal, correction, stereo matching, etc., input the data into the Faster R-CNN model for training, extract target features, and classify and refine in Faster R-CNN Position, complete the training of your own data set, and finally combine binocular and deep learning to detect the target object and its position information in real time.

Figure 5. Overall Process Framework

The reason why the depth learning algorithm is combined with binocular vision is that the core of binocular stereo vision ranging is to calculate the distance between the object and the camera by using the triangle relationship. We all know that the triangle in mathematics is the most stable shape, and has uniqueness. Therefore, the stability of the triangle relationship can make the three-dimensional data more accurate. Monocular vision is the basis of other vision systems, which is relatively simple, but monocular vision has scale uncertainty, and needs a lot of data support, so it can not get absolutely accurate position information. Multi vision not only has its own requirements, but also increases the computational complexity of the algorithm. The common RGBD visual odometer is due to the great influence of light and the limited measurement distance. In

contrast, binocular vision requires relatively low hardware, fast, accurate, flexible and stable, so it is suitable for both indoor and outdoor.

III Results and discussion

The experiment combines 100 degree distortion free binocular camera and Python, Matlab software to achieve, using Matlab to complete the camera calibration, and adjust many times, get the ideal calibration error, the calibration error of this experiment is set at 0.10 pixels.

Figure 6. Binocular Camera

3.1 Experiment and analysis of camera calibration

This experiment uses Zhang Zhengyou calibration method, using Python and binocular camera to collect as many images as possible from different angles, and then load the collected images into the calibration tool of Matlab to run, and calibrate the selected calibration board pictures at the same time. The error of the first calibration is 0.24 pixels. Because there are too many pictures of the calibration board, the calibration error is reduced from 0.24 pixels to 0.16 pixels by three times of manual adjustment and recalibration. The error is 0.12 pixels after the second elimination. After the last elimination of the pictures with error greater than 0.12 pixels, there are still 30 pictures left. After these pictures are calibrated at the same time, the final calibration error is 0.10 pixels.

Figure 7. Calibration Error

It can be seen from the figure that there are still 30 groups of images calibrated at the same time, and some calibrated images can be eliminated again. However, the calibration error has been reduced from 0.24 pixels to 0.10 pixels at the beginning. In addition to ensuring the number of calibration plates, we should also consider the limit of calibration error. We can't blindly pursue the smaller the calibration error, the better. At the same time, we should also combine with the reality. Therefore, considering all factors, it is the best when the error is 0.10 pixels. Finally, fill in the internal and external parameters of the camera, and get the depth map of the object through the binocular camera, as shown in the figure below.

Figure 8. Depth Map

After getting the calibration error, we need to correct, de distort and stereo match the image. Then we can adjust the data to get a better depth map. Before getting the depth map, all the experiments are purely using binocular vision for three-dimensional reconstruction. After we can get the depth map of the object, we can also use the depth map to get the coordinates of the object in the map. In order to get better detection effect, we introduce the depth learning algorithm to continue training.

3.2 Stereo matching experiment

In the stereo matching experiment, we choose the template matching algorithm based on gray level, and on this basis, we compare SAD algorithm, SSD algorithm and NCC algorithm.

Figure 9. Comparison of Three Matching Algorithms

SAD algorithm, SSD algorithm and NCC algorithm all belong to local matching criteria based on gray level. Sad algorithm is similar to SSD algorithm in that it calculates the absolute value or sum of squares of the corresponding pixel differences in two images. NCC algorithm calculates the correlation of the matching regions of two images. As can be seen from figure 9, the final matching results of the three algorithms are almost the same, while sad algorithm is the simplest and the fastest when the template size is determined. SSD algorithm and NCC algorithm are more complex than SAD algorithm.

3.3 Model training loss and analysis

The loss function is the core of the performance of the prediction model, and the appropriate loss function is conducive to the embodiment of the expected results. This experiment calculates and analyzes the loss function in tensorflow target detection API, in which the loss of Faster R-CNN is composed of RPN and Fast R-CNN, and the two parts of loss include classification loss and regression loss.

Figure 10. Loss Function of RPN Network

Figure 10 shows the loss function of RPN network in Faster R-CNN model, which mainly reflects the loss of RPN's bounding box regression and classifier. Because the anchor generated in RPN can only be divided into foreground tag and background tag, the classification loss actually analyzes whether the object is foreground tag or background tag. From the figure, we can see that although the regression and loss functions have individual prominent error values, the overall loss value of regression loss remains between 0.05-0.15, while the overall loss value of classification loss remains between 0.06-0.12.

Figure 11. Loss function of final classification

Figure 11 shows the final classification loss, which actually reflects the loss of Fast R-CNN. Faster R-CNN integrates classification and regression into a network, so the loss function is multitasking. The loss of classification is mainly to classify the detected objects into the loss of each category. The maximum likelihood estimation is used to quantify the accuracy of the classifier. It can be seen from the figure that the classification loss tends to be stable after a certain training, and the loss value is close to 0; although the loss value of regression loss is relatively large, it will be found that after a period of training, there will be a short-term loss close to 0, but the overall trend is downward.

Figure 12. Totalloss and Clone_loss

Figure 12 shows the total loss and clone loss. Since the experiment only trains the model on a single GPU, the two losses are the same, and only the total loss is considered. The total loss is obtained by adding the classification and regression loss of RPN network and the loss of Fast R-CNN network, which reflects the loss of the whole training model. We can see that at the beginning of training, the loss of beating is relatively large, and after reaching a certain training, the loss tends to 0, and has maintained a stable trend.

3.4 Identification result analysis

Object category	Left camera		Right camera	
	X1	Y1	X2	Y2
0	236	239	399	352
1	246	362	423	399
2	489	249	536	286
3	444	302	545	387

Table 2 shows the comparison of the coordinates of the objects captured by the left and right cameras. As is shown in Table 2, the corresponding coordinates of a point in the space in the left image and the right image can be set as (X1, Y1), (X2, Y2) respectively. We know that in order to obtain the three-dimensional information of a point in the image, we need to find the coordinates of the corresponding point in another image, and then we can calculate the three-dimensional coordinates of the point through the internal and external parameters of the camera, correction, matching and other steps.

Figure 13. Three Dimensional Positioning

Figure 13 is the final result of this experiment. Real time detection of indoor objects, according to the binocular camera to detect the location coordinates of objects in three-dimensional space, this is just a screenshot in real-time monitoring. As you can see in the screenshot, each object represents its three-dimensional coordinates relative to the camera, and its category and recognition accuracy are represented by different color boxes. The three-dimensional information in the figure is converted from the data in Table 2. For example, the coordinate of the keyboard relative to the binocular camera in three-dimensional space is (334,330,41), and the unit is centimeter.

IV.Conclusion

In this paper, based on binocular stereo vision, the fusion of deep learning, the use of Faster R-CNN model to complete the data training, and ultimately achieve

three-dimensional positioning of objects. Binocular vision uses "parallax" to obtain the three-dimensional information of the object. Because of the advantages of appropriate accuracy and fast detection speed, it is suitable for places with high performance and environmental requirements, and determines the best calibration error. The SAD algorithm, SSD algorithm and NCC algorithm are compared and analyzed, and the most suitable sad algorithm is selected, which lays the foundation for the subsequent data set training and target detection. Using Faster R-CNN to achieve target detection can make the alignment effect of features and targets better, avoid the repeated extraction and calculation of target features, greatly reduce the amount of calculation, improve the classification accuracy, and the overall performance is about 94.4%, and the system is more stable.

REFERENCES

- [1] Dayu Yan, Wei Song, Xundan Wang, Ziyu Hu. Review of development status of indoor location technology in China[J]. Journal of Navigation and Positioning, 2019, 7(4):5-12.
- [2] Min Shi, Houpan Zhou, Hui Wu, Yiquan Ruan, Ruiqing Qiu. A review of the development and research of indoor positioning technology[J]. Computer Age, 2018, (08):1-4.
- [3] Gaomin Liu, Guofeng Zhang, Longjun Bao. Survey of map construction methods based on monocular vision and simultaneous localization[J]. Application of computer aided design in Chinese Journal of graphics, 2016, 28(6) : 855-868.
- [4] Chong Xu. Research on key technology of indoor positioning based on computer vision[D]. University of Electronic Science and technology, 2020.
- [5] Taixiong Zheng, Shuai Huang, Yongfu Li, Mingchi Feng. Key Techniques for Vision Based 3D Reconstruction: a Review[J]. Journal of Automation, 2018, 46: 631-652.
- [6] Zhejun Tang. Research on 3D reconstruction technology of binocular vision[J]. Information communication, 2020, (06).
- [7] Qiong Wu, Baolong Liu, Ke Wang, Jiang Wang, Hao Lu. The realization and development of binocular stereo vision technology[J]. China New Communication, 2020, 22(02).
- [8] Ming Yang, Haihui Wang, Jun Chen, Ni Wang. The method of improving the accuracy of double target system[J]. Journal of Wuhan University of Technology, 2012(01).
- [9] Xiaojing Cao. Research and implementation of key technologies in binocular stereo vision system [D]. Xi'an University of science and technology, 2019.
- [10] Li Li. Camera Calibration Algorithm Based on OpenCV and Improved Zhang Zhengyou Algorithm[J]. Light Industry Machinery, 2015, 33(04):60-63+68.
- [11] Tan Wang, Leilei Wang, Weiguo Zhang, Xiaotao Duan, Wanli Wang. Design of infrared target system with Zhang Zhengyou calibration method[J]. Optics and Precision Engineering, 2019, 27(08):1828-1835.
- [12] Yan Liu, Tengfei Li. Research of the improvement of Zhang's camera calibration method[J]. Optical Technology, 2014, 40(06):565-570.
- [13] Daixian Zhu, Xiaohua Wang. SLAM algorithm for binocular vision based on Improved SIFT algorithm[J]. Computer Engineering and Application, 2011, 47(14).
- [14] Zhiyuan Wang, Maosen Wang. Distance measurement and positioning system of mobile robot based on binocular vision[J]. Journal of Ordnance Equipment Engineering, 2017(11).
- [15] Shuangquan Li, Guobao Zhang. Binocular Stereo Vision Ranging System Based on ORB Algorithm[J]. Industrial Control Computer, 2017, 30(06):42-44.
- [16] Yabing Lu. Research on real time location and mapping method based on binocular vision[D]. Harbin Institute of Technology, 2018.
- [17] Yifei Zhang, Xinfu Li, Xuedong Tian. SAD stereo matching algorithm based on edge feature fusion[J]. Computer Engineering, 2020,46(02).
- [18] Huijuan Zhang. Research on RGB-D simultaneous location and mapping algorithm in complex environment[D]. University of Chinese Academy of Sciences (Ningbo Institute of materials technology and vehicle supply, Chinese Academy of Sciences), 2019.
- [19] Xiaobo wang, Yong Xie, Xiaori Liu, Binyue Wang. Application and indoor binocular vision positioning[J]. Bulletin of Science and Technology, 2020(8).
- [20] Baocai Yin, Wentong Wang, Lichun Wang. A review of deep learning[J]. Journal of Beijing University of Technology, 2015, 41(01).
- [21] Yang Zhao, Guoliang Liu, Guohui Tian, Yong Luo, Ziren Wang, Wei Zhang, Junwei Li. A survey of vision SLAM Based on deep learning[J]. Robot, 2017(06).
- [22] Shuangquan Li, Guobao Zhang. Binocular Stereo Vision Ranging System Based on ORB Algorithm[J]. Industrial Control Computer, 2017, 30(06):42-44.
- [23] Chenyuan Zhao, Wenxin Li, Qingxi Zhang. Research progress of stereo matching algorithm based on binocular vision[J]. Computer Science and Exploration, 2020(07).
- [24] Qiangwei Jiang, Xingli Gan, Yaning Li. Target recognition and location based on CNN binocular feature point matching [J]. Radio engineering, 2018 (08).
- [25] Hong Tang. Research on grasping method of industrial robot based on deep learning and binocular vision [D]. South China University of technology, 2018 (12).
- [26] Peng Gao. Research on binocular stereo matching algorithm based on

deep learning [D]. Hangzhou University of Electronic Science and technology, 2018.

[27] Chunhui Zhao, Yao Zhou. Ship Target Detection and Recognition Based on Improved Faster RCNN Algorithm[J]. Journal of Shenyang University (Natural Science Edition), 2018, 30(05).

[28] Qi Zhang, Guangdi Hu, Yusheng Li, Xin Zhao. Binocular vision vehicle detection method based on improved Fast-RCNN[J]. Applied Optics, 2018, 39(06) : 832-838.

[29] Songchen Han, Bihao Zhang, Yi Li, Xinmin Tang, Daoyong Fu. Airport surface small target detection algorithm based on improved Faster RCNN[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2019, 51(06).

[30] Linsheng Li, Pingping Zeng. Apple target detection based on improved deep learning framework fast RCNN[J]. Mechanical Design and Research, 2019, 35(05).

[31] Hui Zhang, Yu Du, Shurong Ning, Yonghua Zhang, Shuo Yang, Chen Du. Pedestrian detection method based on Faster RCNN[J]. Transducer and Microsystem Technologies, 2019, 38(02):147-149+153.

[32] Wei Yang, Hongyuan Wang, Ji Zhang, Zhongbao Zhang. An improved vehicle real-time detection algorithm based on Faster-RCNN[J]. Journal of Nanjing University (Natural Science), 2019, 55(02):231-237.

[33] Yanfei Qin. Research on target detection and location technology of picking robot based on binocular vision and deep learning [D]. Beijing Jiaotong University, 2019(01).

[34] Xiangqian Xu, Tao Sun. Pedestrian detection method in traffic scene based on Faster RCNN[J]. Software Guide, 2020,19(04).

[35] Zhijun Cao, Liang Zhang. Fast target detection algorithm based on Faster RCNN[J]. Aerospace Control, 2020,38(04).

[36] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philip Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:4040-4048.

[37] Yan Hu, Zili Shan, Feng Gao. Ship target detection on the sea based on fast RCNN and multiresolution SAR[J]. Radio Engineering, 2018(02).

[38] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(06):1137-1149.

[39] Jipeng Huang, Yinghuan Shi, Yang Gao. Multi-Scale Faster-RCNN Algorithm for Small Object Detection[J]. Journal of Computer Research and Development, 2019, 56(02).

[40] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(06):84-90.

[41] Suofei Zhang, Ye Feng, Xiaofu Wu. Recent advances on object

detection using deep CNNs: an overview[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2019, 39(05):72-80.

[42] Dan Chen, Qingquan Lin. Research on 3D object optimal grasping method based on cascaded Faster RCNN[J]. Journal of Instrumentation, 2019, 40(04):229-237.

Abbreviations

Fast R-CNN : Fast Region with CNN features

Faster R-CNN : Faster Region with CNN features

RPN : Region Proposal Network

ROI pooling: Region of interest pooling

CNN : Convolutional neural network

R-CNN : Regions with CNN features

GPU : Graphic Processing Unit

SAD : Sum of Absolute Differences

SSD ; Sum of Squared Differences

NCC : Normalized Cross correlation

YOLO: you only look once

SPP: spatial pyramid potential

Author information:

Affiliations:

(1)State Key Laboratory of Marine Resources Utilization in South China Sea, Hainan University, Haikou, Hainan, 570228, China

(2) School of Information and Communication Engineering, Hainan University, Haikou, Hainan, 570228, China

Lan Zang

(1)State Key Laboratory of Marine Resources Utilization in South China Sea, Hainan University, Haikou, Hainan, 570228, China

(2)Education Center of MTA, Hainan Tropical Ocean University, Sanya, Hainan, 572022, China

Kun Zhang

Institute of Deep-sea Science and Engineering, Chinese Academy of Science, Sanya, Hainan, 572000, China

Chuan Tian

(1)State Key Laboratory of Marine Resources Utilization in South China Sea, Hainan University, Haikou, Hainan, 570228, China

(2) School of Information and Communication Engineering, Hainan University, Haikou, Hainan, 570228, China

Chong Shen

School of Geography, Nanjing Normal University, Nanjing, Jiangsu, 210023,
China

Bhatti Uzair Aslam

(1)State Key Laboratory of Marine Resources Utilization in South China Sea,
Hainan University, Haikou, Hainan, 570228, China

(2) School of Information and Communication Engineering, Hainan University,
Haikou, Hainan, 570228, China

Ju Huang

Correspondence: Kun Zhang, Chuan Tian, Chong Shen

Authors' contributions

All Authors were involved in manuscript analysis, performance measurement, experimentation, and writing design. All final drafts read and approved by the author.

Funding

This work was supported by the National Natural Science Foundation of China (No. 61861015); 2021 Hainan Province's Major Science and Technology Plan Project "Multimodal Medical Big Data Research and Application Demonstration of Common Key Technologies for Diagnosis and Treatment of Special Diseases" under Grant (No. 202149); 2021 Key Research and Invention Projects of Hainan Province (No. ZDYF2021GXJS032); the Open Project of the State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University under Grant (No. MRUKF2021032).

Availability of data and materials

The data sets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

The authors declare that they have no competing interests.