Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 14, No. 3, 2025

Study on Spatial Localization of Pill Box Gripping Point Based on Binocular Camera

Tao Zhou[1, *]

[1]School of Mechanical Engineering, University of Jinan, Jinan, China
[*]Corresponding author: 17860639221@163.com

Abstract: In order to realize the accurate grasping and localization of the target object by the robotic arm, this paper proposes a three-dimensional spatial coordinate solving method based on binocular vision. By establishing a binocular stereo imaging model, the geometric mapping relationship between spatial points and pixel coordinates is deduced, and a binocular calibration experiment is designed to obtain the camera internal reference and distortion parameters. Aiming at the assembly error of industrial cameras, a stereo correction algorithm is adopted to realize the image polar line constraints, combined with the RGB-D depth image alignment technique to eliminate optical aberrations, and construct a sub-pixel level aligned visual perception system. On this basis, the deep learning pill box detection algorithm is fused to realize the 3D coordinate solution of the grasping target through feature matching. Experiments show that the method can control the localization error within the permissible range of robotic arm grasping.

Keywords: Binocular vision; target location; image alignment.

1. Introduction

The core of robotic arm grasping technology lies in the accurate 3D localization of target objects. By virtue of its bionic stereo sensing characteristics, binocular vision system realizes scene depth reconstruction through parallax computation, which provides a non-contact and cost-effective spatial localization solution for robots [1-3]. In recent years, the fusion of deep learning-based target detection algorithms (e.g., the YOLO series) and stereo matching techniques has driven the rapid development of binocular vision localization methods [4]. Existing research mainly centers on two types of technical paths: one, extracting target features in the left and right views separately through improved target detection networks (e.g., YOLOv5x/YOLOv8), and the other, combining traditional stereo matching algorithms (e.g., SGBM [5]) or deep-learning stereo networks (e.g., RAFT-Stereo [6]) to generate parallax maps, and then extrapolating the depth of the target.

2. Principle of Binocular Vision

The ability of humans to acquire stereoscopic perception of objects through the brain when viewing objects with both eyes is inextricably linked to parallax. The difference between the projected positions of the two retinas is referred to here as binocular disparity. Binocular stereoscopic vision is the conversion of the brain's perception of an image into a geometric computational process, and the ability to perceive an object in three dimensions can be acquired through the brain when two humans observe an object with both eyes, which is inextricably linked to parallax. The difference between the projected positions of the two retinas is referred to here as binocular disparity [99]. Binocular stereo vision is to convert the brain's perception of an image into a geometric computational process by using two cameras to obtain image pairs that mimic the right and left eyes of a human being, respectively, and then obtaining the depth information of the target object through mathematical operations and matrix relation transformation.

A binocular vision model is established as shown in Fig. 1. According to the binocular vision model diagram, the lens optical centers of the left and right cameras can be defined as C_l and C_r, respectively, b denotes the distance between the two optical centers defined as the base distance, the distance from the lens center to the imaging plane is called the focal length denoted by f. p is a point in the real scene, and P_l and P_r are the mappings of the point p on the imaging planes of the left and right cameras, respectively. O_L and O_R are x_2 and x_1, respectively, then define $d = x_1 - x_2$ as the parallax.

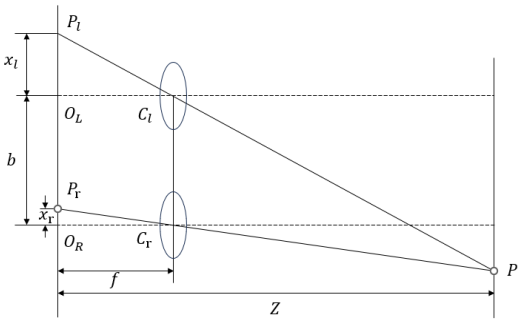


Figure 1. Parallel binocular vision model

As seen in Fig. 1, triangle PP_lP_r and triangle PC_lC_r are similar triangles, and according to the geometric relationship there is:

$$\frac{B + x_l - x_r}{B} = \frac{Z}{Z - f} \tag{1}$$

Then there is:

$$Z - f = \frac{fB}{d} \tag{2}$$

3. Binocular Camera Calibration and Image Alignment

3.1. Camera calibration

To obtain the position information of a point in 3D space by a camera, the prerequisite is to calibrate the camera experimentally to determine the internal parameters of the visual model.

Common camera calibration methods include traditional camera calibration methods [7], camera self-calibration methods [8] [9], and active vision-based calibration methods [10].

Among them, the traditional camera calibration method has higher requirements for known parameters, so although it can obtain more accurate camera parameters, it has some limitations at the application level. Camera self-calibration method compared to the traditional calibration method the method is less accurate, although it can realize real-time calibration, but requires the optical position of the camera to be fixed, so it is not practical. Active vision-based calibration method is more robust, but for the camera movement is unknown or uncontrollable occasions is not applicable.

In summary, several calibration methods of the camera have their own advantages and disadvantages, especially for the harsh requirements of the experimental conditions. In this paper, we comprehensively consider using Zhang Zhengyou camera calibration method [11], which is widely used in camera calibration experiments due to its advantages of simple operation and high accuracy.

The accuracy of the calibration parameters of the camera will directly affect the use of the camera, so it is very important. In this paper, we use the Intel RealSense D435i binocular camera for experiments, which contains an RGB camera, two infrared cameras, and an infrared emitter. In this paper, we use Zhang Zhengyou calibration method to calibrate the parameters of the camera, due to the Zhang Zhengyou calibration method is simple to operate, high precision features, has been widely encapsulated into various types of software can be directly called, this paper with the help of MATLAB toolbox_calib toolbox to calibrate the parameters of the camera.

The specific calibration process is as follows:

(1) Make the calibration board

In this paper, the calibration board is actually checkerboard paper, generated with the help of OpenCV program, as shown in Fig. 2, the specification is 12×9, and the size of a single checkerboard square is 20mm×20mm. in order to avoid the unevenness of the checkerboard paper to the calibration results, the checkerboard paper is printed and pasted on a flat paper board.

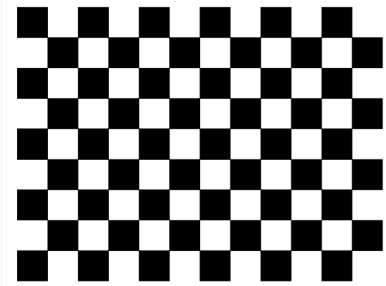
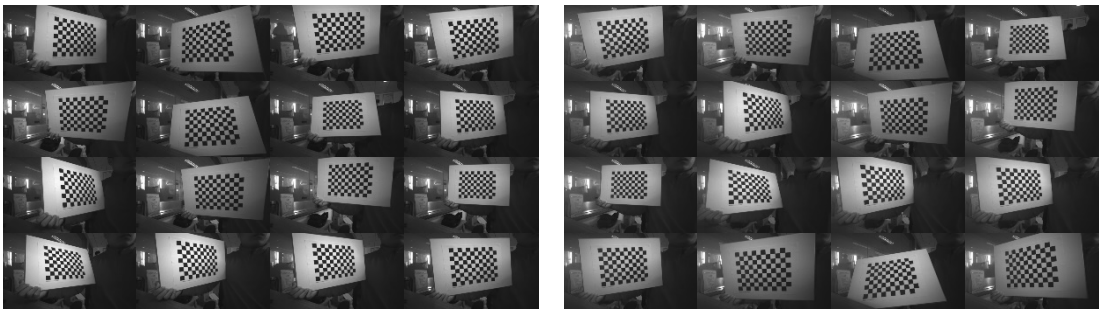


Figure 2. Calibration of checkerboard paper

(2) Image Acquisition

Fix the camera, constantly change the position and angle of the calibration board relative to the camera, respectively, to collect the images of the left and right cameras under different viewpoints, a total of 20 groups of images are collected in this paper. As shown in Fig. 3, the left and right images are part of the images captured by the left and right cameras respectively.



(a) Left camera image (b) Right camera image

Figure 3. Calibration plate acquisition image

(3) Monocular calibration

First of all, the left camera is calibrated with monocular parameters, the left camera image captured in step (2) is imported into Matlab, and the program will carry out the corner point extraction, when the corner point extraction is completed, and after that, the parameters of the left camera are calibrated, and when the calibration is finished, the 3D position simulation of the calibration board can be obtained, as shown in Fig. 4. The calibration process of the right camera is similar to the calibration steps of the left camera.

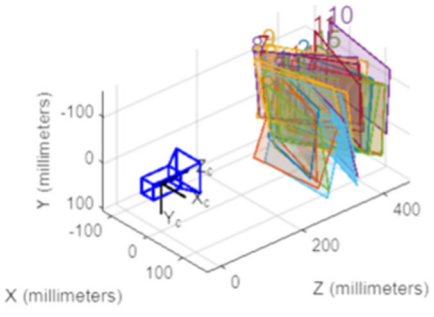


Figure 4. 3D position simulation of the calibration plate

(4) Binocular calibration

Binocular calibration must be done after the left and right cameras complete the single target calibration, the single target calibration results are imported into Matlab, and then complete the binocular calibration through the Matlab toolbox, and then the 3D position of the calibration plate can be obtained, as shown in Figure 5. The calibration results are shown in Table 1.

Table 1. Intel RealSense D435i camera binocular calibration results

Parameter Name	Calibration Results
Left Camera Focus	$[431.1079 \quad 432.3160]$
Right Camera Focus	$[431.3022 \quad 432.8279]$
Left Camera Master Point Coordinate	$[418.3227 \quad 242.4254]$
Right camera principal point coordinates	$[419.0859 \quad 242.0582]$
Left camera radial distortion parameter	$[-0.0096 \quad 0.0509 \quad -0.0789]$
Right camera radial distortion parameter	$[-0.0074 \quad 0.0425 \quad -0.0557]$
Left camera tangential distortion parameter	$[0.0014 \quad -0.0030]$
Right camera tangential distortion parameter	$[0.0015 \quad -0.0028]$
Rotation Vector	$\begin{bmatrix} 1.0000 & -4.2101e-04 & -8.2116e-04 \\ 4.2266e-04 & 1.0000 & 0.0020 \\ 8.2032e-04 & -0.0020 & 1.0000 \end{bmatrix}$
Translation vector	$[-50.0791 \quad -0.0940 \quad 0.4656]$

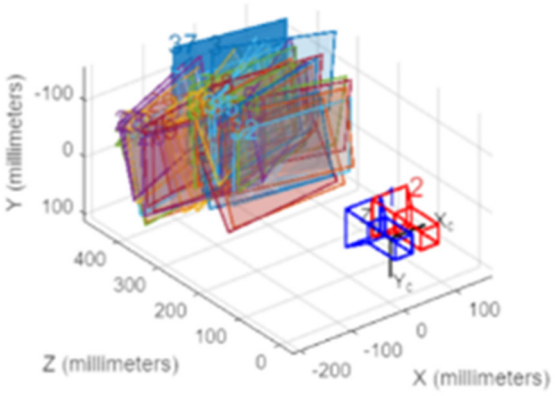


Figure 5. Simulation of 3D position of binocular calibration calibration plate

3.2. Depth image and RGB image alignment

D435i binocular camera has two infrared cameras and one RGB camera, the left and right two infrared cameras can obtain the depth image containing distance information, which can be directly used for localization, and the RGB camera can obtain the RGB image containing the image color, texture and other information, in order to ensure that the three-dimensional coordinates of the acquired target object and the two-dimensional image of the correspondence is accurate, it is necessary to carry out the depth image and RGB image alignment. Alignment of depth image and RGB image.

To ensure the accurate correspondence between the 3D coordinates of the target object and the 2D image, it is necessary to align the depth image with the RGB image. The transformation matrix L between pixel coordinate systems is obtained as:

$$L = \begin{bmatrix} 1 & 0.0008 & -0.0023 & -0.0149 \\ -0.0008 & 1 & -0.0020 & 0 \\ 0.0023 & 0.0020 & 1 & -0.0005 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

After the conversion matrix L is obtained through the image alignment experiment, the depth information of the target points in the RGB image can be corresponded to the coordinates of the depth image one by one, so as to realize the alignment of the RGB image with the depth image. As shown in Fig. 6, the RGB image captured by the camera and the depth image after alignment are shown, respectively.

(a) RGB image (b) Depth image after alignment

Figure 6. Alignment image of D435i camera

4. Acquisition of Three-dimensional Coordinates of the Pill Box

4.1. Acquisition of drug box spatial coordinates

By observing the drug box ROI recognition effect graph analysis, choose the drug box ROI x direction center and y direction 1/3 as the three-dimensional spatial coordinates of the drug box to absorb the point. As shown in Equation 3 for the parameter structure of the pill box enclosing box:

$$box = (x, y, w, h, lab_{\_id}, prob) \tag{3}$$

Where, x and y are the pixel coordinates of the upper left corner of the wraparound frame in the X and Y directions respectively, w and h are the width and height of the wraparound frame respectively, $lab_{\_id}$ is the category of the detection result to be displayed, and prob is the confidence level of the detection result. The parameter information are the values under the pixel coordinate system of the image, and the upper left corner of the image is defined as the origin of the pixel coordinate system.

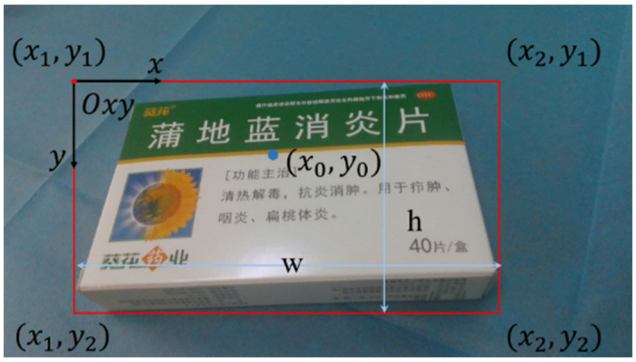


Figure 7. Schematic diagram of the location of the center point of the ROI of the pill box

So far, the pixel coordinate value of the center point (x_0, y_0) of the enclosing box is obtained by calculation as shown in Fig. 7:

$$x_0 = x_1 + w/2 \tag{4}$$

$$y_0 = y_1 + h/3 \tag{5}$$

After obtaining the pixel coordinates of the center point of the ROI of the pill box, the pixel coordinates are brought into the RGB image and transformed by the camera coordinate system so as to obtain the 3D coordinate parameters $[X_c, Y_c, Z_c]$ of the center point of the ROI of the pill box under the camera coordinate system in unit of m. The effect of real-time acquisition of the 3D coordinates of the pill box is shown in Fig. 8.

Figure 8. The effect of obtaining the 3D coordinates of the pill box

5. Summary

In order to realize the precise grasping operation of the robotic arm on the target object, this study carries out a systematic research around the three-dimensional spatial localization problem of the target object. Firstly, a mathematical model of binocular imaging is established based on binocular stereo vision theory, and the geometric mapping relationship between the spatial coordinate system and the image pixel coordinate system is deduced. The internal reference matrix and distortion coefficient of the camera are obtained by building a high-precision calibration platform. On this basis, this study innovatively integrates the depth image alignment technique to realize the alignment of RGB images with depth information. The experimental results show that combined with the improved YOLOv8 pill box detection algorithm, the system can accurately acquire the 3D coordinates of the pill box gripping center point.

References

[1] CONG Y, CHEN R, MA B, et al. A comprehensive study of 3-D vision based robot manipulation [J]. IEEE Transactions on Cybernetics, 2023, 53(3): 1682-1698.

[2] WANG C, CUI X, ZHAO S, et al. The application of deep learning in stereo matching and disparity estimation: a bibliometric review [J]. Expert Systems with Applications, 2024, 238: 122006.

[3] POGGI M, TOSI F, BATSOS K, et al. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5314-5334.

[4] ZOU Z, CHEN K, SHI Z, et al. Object detection in 20 years: a survey [J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[5] HIRSCHMULLER H. Stereo processing by semiglobal matching and mutual information [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 30(2): 328-341.

[6] LIPSON L, TEED Z, DENG J. Raft-stereo: multilevel recurrent field transforms for stereo matching [C]// Proceedings of the 2021 International Conference on 3D Vision. London: IEEE, 2021: 218-227.

[7] ROGER,Y. Tsai. An Efficient and Accurate Camera Calibration Technique For 3D Machine Vision[C]// Conference on IEEE Computer Society. 1986.

[8] O. D. Faugeras, Q. -t. Luong, S. J. Maybank. Camera Self-Calibration: Theory and Experiments[J]. proc of eccv, 1992, 588(12):321-334.

[9] Maybank S J , Faugeras O D . A theory of self-calibration of a moving camera[J]. International Journal of Computer Vision, 1992, 8(2):123-151.

[10] Ma S D . A self-calibration technique for active vision systems[J]. IEEE Transactions on Robotics & Automation, 1996, 12(1):114-120.

[11] Zhang, Zhengyou. A Flexible New Technique for Camera Calibration.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000.