Calculating Distance to Tomato Using Stereo Vision for Automatic Harvesting

Z. Buribayev[1,2][0000−0002−3486−227X],
T. Merembayev[2,3][0000−0001−8185−235X],
Y. Amirgaliyev[1,2][0000−0002−6528−0619],
T. Miyachi[4], and
A. Yeleussinov[1,2][0000−0002−0425−6527]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2] Institute of Information and Computational Technologies, Almaty, Kazakhstan
[3] International IT University, Almaty, Kazakhstan
[4] School of Information Science and Technology, Tokai University, Hiratsuka, Japan

Abstract An accurate estimate of the distance between the robot and equipment is essential for the application of robots. Sensors such as laser and sonar are mainly used to calculate distances. The research considers a pair of Web cameras for distance measurements To solve this problem, the stereo vision method was proposed. A comparison of the accuracy of the calculations was done for 3 zones: left, right and center. This separation allows us to identify possible distortion or camera failure at the edge of the image. As a result of research, the accuracy of 71.939% was achieved. Also defined parameters which can help to improve the precision of object coordinates in 3D.

Keywords: Stereo vision, Robots, Pattern recognition, Disparity map, Depth map;

INTRODUCTION

Using computer vision technology is widely implemented in various fields and also has commercial success. Due to the success of the technology, scientific research in this area is of great interest to scientists. In the direction of computer vision, there are two large areas of research: image processing and pattern recognition. This paper discussed the problems of image processing, this is the first step for future research in the area of pattern recognition.

Computer vision devices can be divided into two categories:

- 2D - obtaining an image with no image depth.
- 3D - image acquisition with which you can calculate Z values.

2D images.

Image data is obtained from a mono camera and the result is a 'flat' image. Often, these cameras are installed on devices that perform dynamic movements and decompose 2D images in time.

3D images.

Image data is achieved by using a stereo pair of images. The device that takes photographs consists of two identical cameras.

RELATED WORK

Normal monocular visuals suffer from scale bias. Pioneering researchers [1,2,6,7] show that this problem can be mitigated by learning from 2D flow functions. Inspired by RGBD-SLAM, relative conversion can be estimated directly from the solution to the PnP problem when depth is specified.

To solve this problem, the authors used deep learning algorithms. CNN-SLAM [8] is a precursor for learning depth prediction training with monocular SLAM to create an accurate dense 3D map.

The research [6] considers the solution to the problem of calculating three-dimensional (3D) coordinates for a material point. For this purpose, two flat images (stereo pair) are used, which correspond to the left and right view points of the 3D scene. A stereo pair is obtained using two cameras with parallel optical axes. A series of experimental studies were conducted to verify theoretical results. During these experiments, aminor discrepancy was caused by the spatial distortion of the camera (distortion) in the optical system and its discrepancy. When using a high-quality stereoscopic system, the existing discrepancy calculation allows applying this method to a wide range of practical problems.

In [7], the errors of calculating the distance to an object using a stereoscopic system are analyzed. It was found that the percentage error in calculating the distance is inversely proportional to the number of pixels used when shifting between two images, and is directly proportional to the distance to the object.

The application of stereo image processing is used for geology interpretation based on the image log (image of downhole). It helps to save time for core sample interpretation [8].

Based on the review of the work, we highlight the following conclusions:

- Determination of distance using a mono camera does not have widespread use and research. Although the use of monocamera should reduce the cost of production of technical devices.
- Reducing the error of distance determination, research in the field of stereo vision is not sufficiently studied and has prospects for a more detailed study.

In this research, we attempted to solve the problems obtained from the literature reviews. This research describes an experiment with a stereo camera, calculating the error and identifying the parameters that affected this error.

METHODS

The paper describes a method of training artificial intelligence to recognize the distance to an object using the modern computer vision detector YOLO, using triangulation.

When shooting objects, a monocamera was used, and to calculate the distance between the camera and the object, the image is transmitted from the camera to the object classifier, based on the modern computer vision object detector YOLO (You Only Look Once). YOLO is a fast and accurate object detector based on the Convolution Neural Network (CNN) [9]. The output is the bounding borders of the detected objects in the image and the class labels of the detected objects. First of all, objects are classified by certain parameters, in our case by color. After classification, parameters are processed to calculate the coordinates of the bounding frames. Based on the coordinates of the bounding frame, the distance from the object to the camera is calculated. The block diagram of the proposed system is illustrated in Figure 1, an example of estimating tomato distances. To make research work realistics and apply the results in any conditions, we used the Tomato DataSet, obtained when shooting objects in a greenhouse complex without the participation of professional shooting mechanisms (prof. photographer, special lighting, etc.). Tomato DataSet was collected in the greenhouse complex ”BRB APK”located in Almaty city, Kazakhstan.

The measurement of the distance between the robot and the object is necessary to control the actions of the robot, such as capturing the object or even avoiding obstacles [10,11]. There are many methods for estimating distance, such as ultrasound, laser, and (video, photo) cameras. A technique based on a video camera has the advantage of its low cost, so in this paper, we will describe a method for measuring the distance between a stereo camera and an object (tomato).

The process of calculating the distance includes five steps: calibrating a stereo camera, calculating a disparity map, calculating a depth map, calculating 3D coordinates of an object in the real world, calculating the distance between a stereo camera and an object.

The stage of calibrating a stereo camera includes the processes of straightening the image and filtering the received image. Calibration of a stereo camera is the process of obtaining internal parameters and external parameters. The internal parameters are the result of the distortion of the lens, while the external parameters depend on the modeling of geometric relationships between the two cameras. For this process, we simultaneously shot a chessboard from two webcams. To increase the quality of the calibration, we took 65 photos Figure 2.

During the calibration process, a checkerboard corners are searched for in all images shot by the left and right cameras simultaneously. The position of the corners for each image is then stored in the image vector, and the object points for the 3d scene are saved in another vector. The image correction process is then carried out using these values. To successfully calibrate the stereo camera, 65 pairs of chessboard images were used. After the completion of the process, we got such data as: camera matrix, distortion coefficients, rotation and displacement vectors

To calculate the stereo camera mismatch map, we used the OpenCV libraries Stereo SGBM algorithm. The Stereo SGBM algorithm is based on the idea of pixel-by-pixel matching and the subsequent application of global twodimensional constraints [13]. The task of calculating parallax in SGBM is formulated as the problem of minimizing the similarity criterion:

$$S(p,d) = \sum_{r} L_r(p,d) \tag{1}$$

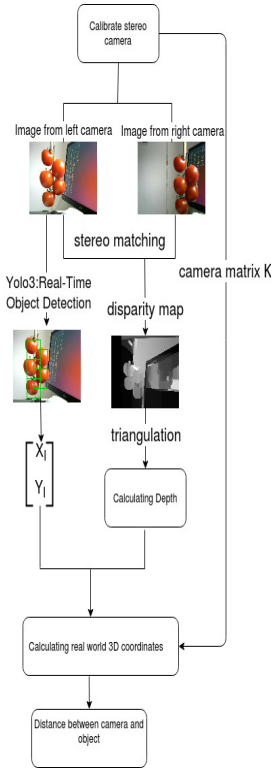

Figure 1. The block diagram of the proposed system.

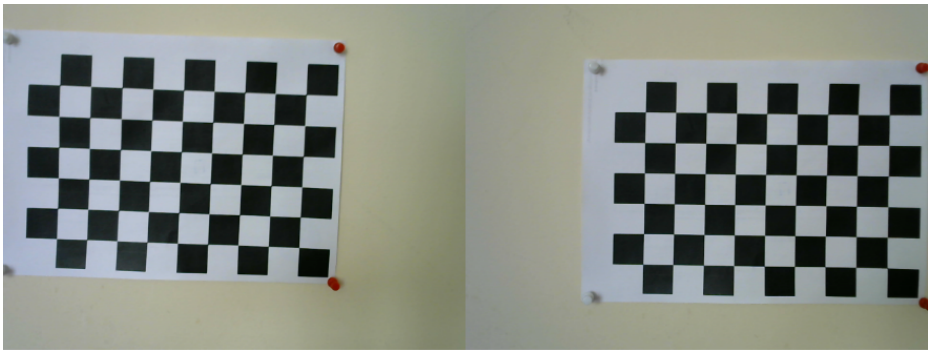

Figure 2. Example of calibrating a stereo camera.

where is the p-pixel of the first image, d is the horizontal parallax of this pixel, and L is a quantity characterizing the path that has been traveled in the r direction. The mismatch map for the base image is calculated, as in the usual methods of local matching, by choosing for each pixel p such an offset d that meets the least similarity criterion, i.e. $min_S(p, d)$.

The pixel disparity value is often interpreted as the inverse distance to the observed objects. In other words, the mismatch is inversely proportional to depth. Therefore, detecting non-conformities is important for building a depth map.

$$Depth = \frac{f * b}{Disparity} \tag{2}$$

where f is the focal length of the camera, b is the distance between the cameras, Disparity is the value of the mismatch map. 3D coordinates of the object in the real world are calculated according to formula 3.

$$\begin{aligned} X &= \frac{Depth * (x_{pixel} - c_x)}{f_x} \\ Y &= \frac{Depth * (y_{pixel} - c_y)}{f_y} \\ Z &= Depth \end{aligned} \tag{3}$$

where $Depth$ is the pixel depth value, fx, fy are the focal lengths expressed in pixel units, cx, cy is the main point that is usually located in the center of the image, X pixel, Y pixel are 2D coordinates in pixels.

After completing the above steps, to calculate the distance between the stereo camera and the object, we used the Euclidean distance according to 4.

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{4}$$

where, $x1$, $y1$, $z1$ are the coordinates of the stereo camera, $x2$, $y2$, $z2$ are the coordinates of the object.

In this study, we used a Logitech HD Webcam C270 webcam with the following Table 2 specifications.

Table 1. CAMERA SPECIFICATION.

N	Type	Description
1	The matrix	Resolution: 1280 x 720 pixels
2	Camera resolution	Up to 3 megapixels with software processing
3	Max Resolution	720p/30fps
4	FoV(°)	60
5	Focus type	fixed focus

RESULTS AND FUTURE WORK

The purpose of the experiment is to calculate the distance from the stereo camera to the desired object (tomato). A detected object is tracked using the bounding box shown in Fig 3. When testing the average detection in the frame takes 69 ms, it is 15-16 FPS (frames per second).

Table **??** displays the experimental results. As you can see, the calculated distance data does not quite correspond to the real data. This problem is related to the calculation of the disparity value and it depends on four factors such as:

1. Pulsed lighting
2. Short exposure
3. Uncoordinated shutters
4. Large low contrast areas in the scene

Figure 3. An example of detecting a tomato.

Based on the data obtained, errors between the real distance and that calculated using the camera were calculated. The following metrics were used for this calculation: R2 and Mean squared Table 2.

A comparison of the accuracy of distance calculation is divided into 3 zones, left, right, center. This separation allows us to identify possible distortion or camera failure at the edge of the image. Based on the results of Table 2, metric accuracy was obtained. The general accuracy of the distance was also calculated for all measurements - General.

In Fig. 4 shows the tabular results of TABLE 3 in graphical form.

Based on the developed algorithm, it is possible to improve the accuracy and speed of distance calculation by using parallel computing algorithms. Also, given the results, it is possible to apply this method to the TakoBot robot [13].

Table 2. Actual vs measured distance.

Real distance (cm)	Estimate distance Right (cm)	Estimate distance Left (cm)	Estimate distance center (cm)
10			8.19
15			10.48
20	12.01		14.33
25	14.49	13.68	13.47
30	15.6	16.88	16.67
35	17.78	18.85	18.91
40	19.83	21	21.68
45	22.51	22.22	24.81
50	24.81	23.71	28.37
55	29.47	25.16	29.63
60	29.47	29.31	33.34
65	29.97	32.33	35.8
70	31.3		37.04
75			39.22
80			

Table 3. DISTANCE ACCURACY METRIC.

Zone of measurement	R2 metrics	MSE metrics
Center	0.869	98.314
Left	0.946	32.437
Right	0.879	81.404
General	0.894	71.939

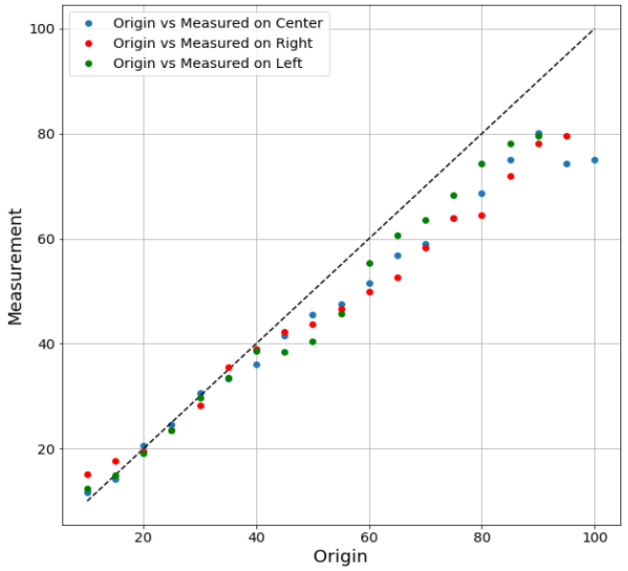

Figure 4. Graphical display comparing the results of known distances and measured.

CONCLUSION

This experiment is conducted with a low-resolution camera to verify whether the proposed algorithm is and the measurement accuracy can be improved using a highresolution camera. Image noise is the main unavoidable cause of errors during the image acquisition phase. Such errors can occur when finding the exact point of contact with the object on the ground. Another potential cause of the error is image change.

Also, the accuracy of the measurement is illuminated, whether it is natural or artificial lighting. Testing this hypothesis requires additional research.

The study conducted an experiment to identify parameters that would improve the accuracy of calculating the distance to the object using one fixed camera, even if the surface of the object is not parallel to the camera and the object is not limited by the vertical movement of the optical axis.

ACKNOWLEDGMENT

This work is supported by a grant from the Ministry of Education and Science of the Republic of Kazakhstan within the framework of the Project ”AP05132648 - Creating verbal and interactive robots based on advanced voice and mobile technologies“. Conflicts of Interest: The authors declare no conflict of interest.

References

1. Costante, Gabriele, and Thomas Alessandro Ciarfuglia.: LS-VO: Learning dense optical subspace for robust visual odometry estimation. IEEE Robotics and Automation Letters **3**(3), 1735–1742 (2018)
2. Clark, Ronald, et al.: Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. Thirty-First AAAI Conference on Artificial Intelligence. 2017.
3. Muller, Peter, and Andreas Savakis.: Flowdometry: An optical flow and deep learning based approach to visual odometry. In: IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, (2017)
4. Wang, Sen, et al.: Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017.
5. Tateno, Keisuke, et al.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017).
6. Mussabayev, R. R., Kalimoldayev, M. N., Amirgaliyev, Y. N., Tairova, A. T., Mussabayev, T. R.: Calculation of 3D Coordinates of a Point on the Basis of a Stereoscopic System. Open Engineering. **8**(1), 109–117 (2018)
7. Marengoni, Mauricio, and Denise Stringhini.: High level computer vision using opencv."In: 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials. IEEE, (2011).
8. Merembayev T., Yunussov R. and Amirgaliyev Y.: Machine Learning Algorithms for Classification Geology Data from Well Logging. 14th International Conference on Electronics Computer and Computation (ICECCO) pp. 206–212, IEEE, (2018)
9. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi.: You Only Look Once:Unified, Real-Time Object Detection. Computer Vision and Pattern Recognition, (2016)
10. A. Yeshmukhametov, Z. Buribayev, Y. Amirgaliyev, B. Amirgaliyev, K. Latuta.: Bio-inspired a novel continuum robot arm with variable backbone design: Modelling and validation. Journal of Theoretical and Applied Information Technology. **97**(19), 5036-5047 (2019)
11. Yeleussinov, T. Islamgozhayev, M. Satymbekov and A. Kozhagul.: CVCER: Robot to Learn Basics of Computer Vision and Cryptography. IOP Conference Series: Materials Science and Engineering, **417**(1), (2018)
12. StereoSGBM, `http://docs.opencv.org`. Last accessed 11 Sep 2020
13. A.Yeshmukhametov, K.Koganezawa, A.Seidakhmet, Y. Yamamoto.: A Novel Passive Pretension Mechanism for WireDriven Discrete Continuum Manipulators. Proceedings of the 2020 IEEE/SICE International Symposium on System Integration Honolulu, Hawaii, USA, (2020)