Object distance estimation using a monovision camera

Efemenah Endurance Idohen, Afolayan Matthew Olatunde, Umar Ali Umar
Department of Mechanical Engineering, Faculty of Engineering, Ahmadu Bello University, Zaria, Nigeria

Article Info

Article history:

Received May 10, 2023
Revised Jun 4, 2023
Accepted Jul 3, 2023

Keywords:

Depth estimation
Distance estimation
Distance measurement
Monovision camera
Object
Pixy2 camera

ABSTRACT

In computer vision, most monovision cameras used for estimating the position of an object only estimate the 2D information of the object without the depth information. Estimating the depth information, which is the distance between the target object and the camera is quite challenging but, in this paper, a less computationally intensive method was used to estimate the object's distance to complete the 3D information needed to determine the object's location in cartesian space. In this method, the object was positioned in front of the camera at a sequential distance and was measured directly. The distances measured in the experiment with a set of training data obtained from the image were fitted into a curve using the least-square framework to derive a non-linear function that was used for estimating the object's distance also known as the z-coordinate. The result from the experiment showed that there was an average error of 1.33 mm between the actual distance and the estimated distance of the object. Hence, this method can be applied in many robotic and autonomous systems applications.

Corresponding Author:

Efemenah E. Idohen
Department of Mechanical Engineering, Ahmadu Bello University
P.M.B 1045, Zaria, Nigeria
Email: efemenah@gmail.com

1. INTRODUCTION

Estimating the depth information of an object from its pose in an environment is an essential part of computer vision but with monocular cameras, it is quite difficult to estimate the object's depth. Generally, monocular cameras acquire only 2D information about an object from a scene by virtue of perspective transformation which results in a loss of depth information [1], [2]. Therefore, obtaining the depth information to have complete 3D information about the object's pose can be useful in many robotic applications such as pose estimation, picking and placing, and mapping. Traditional methods such as the use of Bluetooth, laser, ultrasonic and IR sensors have been used in the past to estimate the object's distance [3]–[5] but with the advent of vision sensors, stereo vision and monocular vision are the only two predominant methods used for estimating the object's distance in image-based visual servoing. The stereo vision, which is also known as the computer-based passive approach uses two cameras in the form of binocular structure or human eyes to estimate the depth information of the object [6]–[8]. This can be achieved by placing two cameras horizontally apart and at equal distances from their center points to capture 2D images of the object in their views [9]. Due to the distance separating the two cameras, the captured images are known as disparity images and are used for computing the depth information at the point where the field of view of the two cameras intersects. The stereo vision method is highly accurate but requires a large number of images to be processed in order to achieve precision. It also requires many complex computations due to the large number of images used hence, it is computationally intensive. This method is also expensive to implement because it requires the use of two cameras. In contrast to the stereo vision method, the monocular vision method involves the use of a single camera to estimate the object's distance based on the reference points of the camera's field of view [10]. This method is fairly accurate but not computationally intensive because it requires only a few image registrations that enable the computer to process the images faster. Thus, this type of method can effectively reduce the system workload and save the computer a longer processing time [11]. The monocular method used for visual servoing purposes is cheap and has low handling complexity due to the use of only one camera.

2. RELATED WORK

Object distance measurement plays a vital role in the acquisition of objects' depth information that complements the classic 2D visual perception used for robotic and autonomous systems applications. However, brief literature on distance estimation is presented in this section. Zhou *et al.* [12] used a monocular vision method to find the position and orientation of the object at a distance of 5 m. The relative translation and rotation values of X, Y, and Z directions were obtained through an unconstrained linear equation of rotation and translation matrix R, T and were computed using the inverse least-square method. Krishnan *et al.* [13] proposed a method of complex log mapping to measure the distance between the camera and the object's surface with an arbitrary pattern. The method is based on the use of two images taken at two different camera positions that are known while moving the camera along its optical axis. The distance of the object to the camera is therefore estimated by computing the ratio between the sizes of the object projected on the two images.

Chang *et al.* [14] proposed an efficient neural network method for achieving self-localization by a humanoid robot. Yang and Cao [15] also proposed a 6D pose estimation of an object using the Levenberg-Marquardt algorithm to refine the result of the decomposed homography matrix. Zhang *et al.* [16] proposed a method of estimating the localization of an object that is based on perspective transformation. Their method was presented in three stages. The first stage dealt with the calibration of the camera to calibrate the intrinsic parameters. The second constituted a model for computing the object's distance through perspective transformation by mapping the 3D points in the real world to the 2D image of a pinhole camera. The third stage, which is the measurement of the absolute distance between the camera and the target object, was achieved through the geometry formed from the perspective projections.

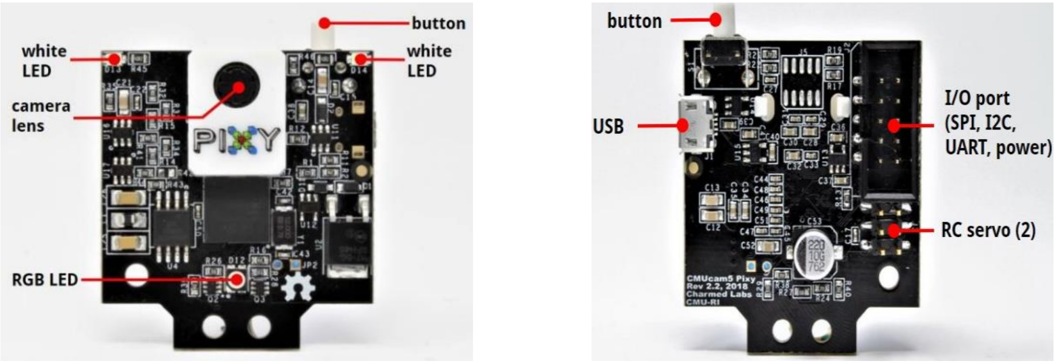
Muslikhin *et al.* [17] used a machine learning algorithm to classify the positions of the object in the image of the mono camera and then used the k-nearest neighbors (k-NN) approach to find the nearest point of the centroids to the closest class. Bui *et al.* [18] proposed the use of a single camera with a triangulation method to measure the distance of an object indirectly. The method is such that the distance to the object is determined based on one known angle and two sides of a triangle. Zheng *et al.* [19] presented a method of measuring an object's distance by a monocular vision camera on a mobile robot. However, the distance between the mobile robot and the target object was determined based on the sub-pixel image processing, mapping, and path planning method. Zhu and Fang [20] initially proposed to address the distance estimation problem with a deep-learning-based method by predicting directly the distance of a given object on red, green, and blue (RGB) images without the use of intrinsic parameters of the camera. They further enhanced the model with a key point regressor in which a projection loss was defined to estimate the distance of objects close to the monocular camera while facilitating the training and evaluation tasks with extended KITTI and nuScenes (mini) datasets of specified objects' distances.

Vajgl *et al.* [21] presented a Dist-YOLO method that is based on YOLO architecture in which the original loss function is updated to estimate the absolute distance of an object using the information from the monocular camera. Most of the methods used for estimating the object's distance in the literature are computationally intensive but, in this paper, a monovision camera was used to obtain a set of image-based data with the measured distances of the object and was computed by using a curve fitting technique to derive a non-linear function for estimating the object's distance.

3. METHOD

To determine the distance of the object from the camera, which is the depth information, a single Pixy2 camera was used in this study. The Pixy2 camera is a vision sensor with an embedded image processor that can process captured RGB images and segment them to recognize objects of different colors while using its built-in color-based filtering algorithm called the color-connected components (CCC). As it has the capability of tracking up to seven different colors, which are red, blue, green, yellow, orange, cyan, and violet, it also has the functionality of tracking the object's position in the image in two dimensions The front and back of the views of the Pixy2 camera is shown in Figure 1.

Though the Pixy2 camera can perform other functions such as line tracking and barcode reading [22], in this study, it will be used to train a specific object with a single color positioned at a sequential distance from the camera to acquire a dataset for determining the object's distance.



Front view

Back view

Figure 1. The Pixy2 camera

3.1. Camera set-up

To train the Pixy2 camera to acquire the visual information of the object found in its field of view, the vision sensor needs to be installed in a position where the target object will be visible to the camera in order to avoid occlusion. So, the eye-in-hand configuration was used in this paper. The eye-in-hand configuration is a posture the camera takes when mounted on a manipulator and it can either be after or before the wrist of the robotic arm [23], [24]. Figure 2 shows the Pixy2 camera mounted on the robot manipulator that is used for a pick and place purpose.

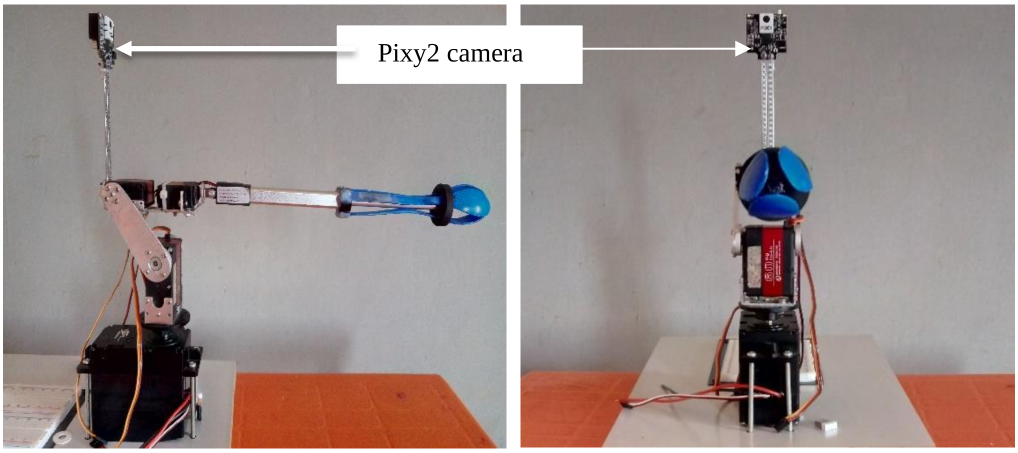


Figure 2. Pixy2 camera mounted on the elbow joint of the manipulator

3.2. Distance measurement using a single Pixy2 camera

To measure the distance of the object using a single Pixy2 camera, a set of training data that can be used for estimating the object's distance was generated first from the experiment. However, in this method, a ripe tomato which is completely red was used as the target object in the experiment and was trained to be recognized by the Pixy2 camera using its PixyMon software. The ripe tomato was simultaneously positioned at a horizontal distance between 430 and 580 mm in front of the robotic arm in the real world; and a vertical distance between 0 and 207 mm of the camera's image height. The horizontal and vertical distance parameters used in training the object were based on the manipulator's length (580 mm) and the entire image height (207 mm) of the camera. The object (ripe tomato) was placed sequentially in the camera's field of view (FOV) as shown in Figure 3.

However, on placing the ripe tomato sequentially in the camera's FOV, the respective distances of the ripe tomato from the camera's lens were measured using a measuring tape with an accuracy of $\pm 0.5\ mm$. Therefore, to generate training data, the actual distances measured were recorded alongside the image data generated by the Pixy2 camera. The image data consists of the two coordinates (x, y), the width and height of the ripe tomato to determine the area of the bounding box as shown in Figure 4. These were estimated by the

Pixy2 camera based on its image processing and object tracking capabilities when the ripe tomato was placed sequentially within the specified horizontal and vertical distance parameters. The training data obtained from the experiment by placing the ripe tomato in sequential positions relative to the camera's reference position is given in Table 1.

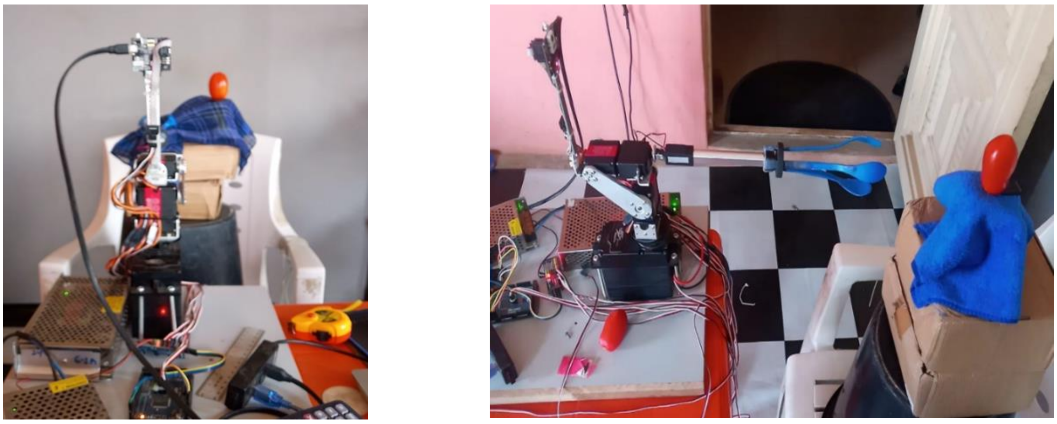
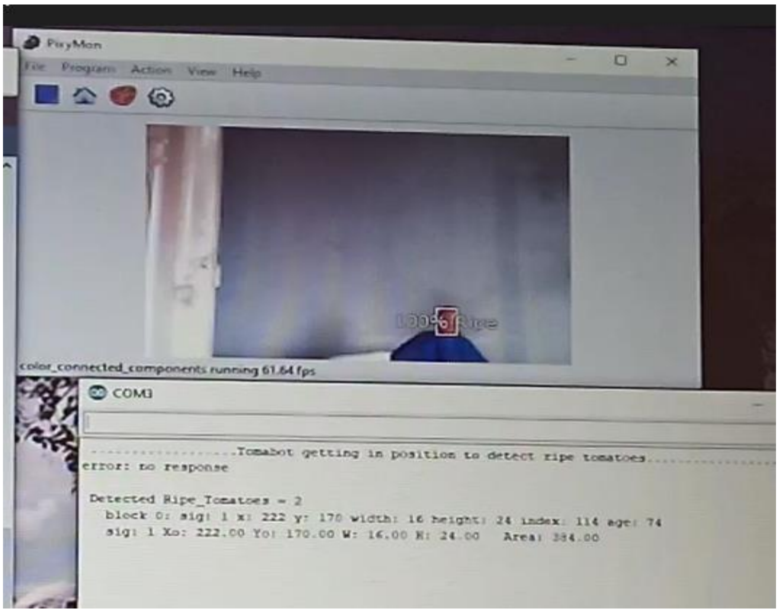


Figure 3. A ripe tomato (object) placed sequentially in the camera's field of view



```
color_connected_components running 61.64 fps
```

```
..................Tomabot getting in position to detect ripe tomatoes.................
error: no response

 Detected Ripe_Tomatoes = 2
   block 0: sig: 1 x: 222 y: 170 width: 16 height: 24 index: 114 age: 74
   sig: 1 Xo: 222.00 Yo: 170.00 W: 16.00 H: 24.00    Area: 384.00
```

Figure 4. A captured ripe tomato bounded by a box in the image to obtain the trained image data for the computation of the object's distance

Table 1. Data obtained from training the Pixy2 camera to estimate the positions of the ripe tomato when placed sequentially in the camera's field of view

Trial	$\boldsymbol{X_c}$(mm)	$\boldsymbol{Y_c}$ (mm)	Width (mm)	Height (mm)	Area (mm^2)	Actual Distance (mm)
1	274	97	24	22	528	430
2	216	115	24	19	456	438
3	189	125	22	20	440	463
4	168	134	20	19	380	490
5	140	142	20	18	360	502
6	121	151	18	18	324	530
7	107	166	18	17	306	545
8	64	182	16	17	272	550
9	28	198	16	12	192	575

However, to determine the object's distance, which is the *z-coordinate* of the ripe tomato irrespective of its pose in the camera's FOV, the least-square method which takes the best-fit curve from a given dataset with a minimal sum of deviations [25] was employed to obtain the relationship between the area of the bounding box and the actual distance obtained from the training data in Table 1. The curve-fitting plot produced a non-linear relationship between the actual distance and the area of the bounding box in Figure 5.

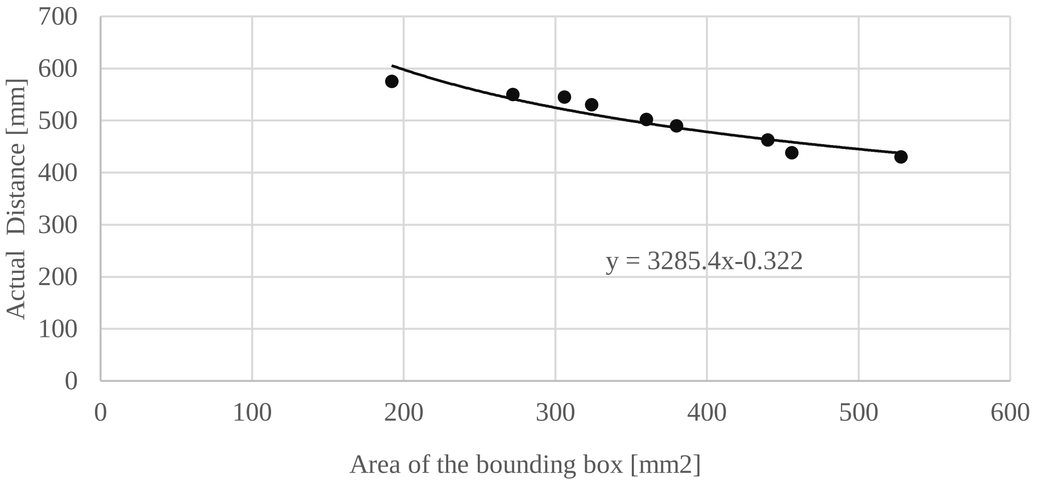


Figure 5. The graph of the actual distance against the area of the bounding box

The non-linear function obtained from the graph is presented in (1),

$$y = 3285.4x^{-0.322} \quad (1)$$

where y is the actual distance and x is the area of the bounding box. Hence, the distance is as in (2).

$$\text{Distance} = 3285.4(\text{Area})^{-0.322} \quad (2)$$

However, the relationship between the actual distance and the area of the bounding box variable in (2) was used to estimate the distance of the object from the camera.

4. RESULTS AND DISCUSSION

To estimate the object's distance, the distance-area relationship in (2) was used to estimate the distance of the ripe tomato from the Pixy2 camera using the area of the bounding box and the actual distance data in Table 1. Hence, the result was validated by determining the average error of the difference between the actual distance and the estimated distance. It can be seen from Table 2 that the slight deviation in the estimated distance resulted in an average error of 1.33 mm. Also, both estimated and actual distances were compared graphically as shown in Figure 6.

Table 2. Result of the estimated distance and the average error

Trial	Area (mm^2)	Actual Distance (mm)	Estimated Distance (mm)	Error (mm)
1	528	430	436	-6
2	456	438	458	-20
3	440	463	463	0
4	380	490	485	5
5	360	502	494	8
6	324	530	511	19
7	306	545	520	25
8	272	550	540	10
9	192	575	604	-29
		Average error		1.33

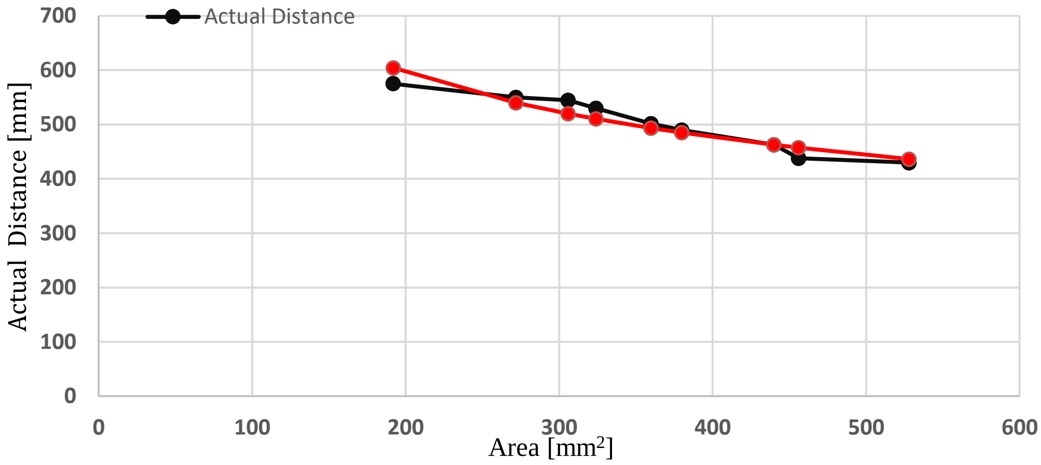


Figure 6. Comparison of the estimated distance and actual distance of the object

5. CONCLUSION

The low-cost monovision camera and the least-square method used in this paper can estimate the distance of the object from the camera irrespective of its pose in the camera's field of view under varying light conditions. The result from the experiment shows that the average error from the estimated object's distance is 1.33 mm. However, since this method is capable of complementing the 2D information that can be used for determining the object's location in cartesian space, therefore, it can be applied to many robotic and autonomous systems applications.

REFERENCES

[1] J. Wu, D. Yin, J. Chen, Y. Wu, H. Si, and K. Lin, "A survey on monocular 3D object detection algorithms based on deep learning," *Journal of Physics: Conference Series*, vol. 1518, no. 1, Apr. 2020, doi: 10.1088/1742-6596/1518/1/012049.

[2] R. Goroshin, "Obstacle detection using a monocular camera," Thesis, Georgia Institute of Technology, 2008.

[3] M. S. Bargh and R. de Groote, "Indoor localization based on response rate of Bluetooth inquiries," in *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*, Sep. 2008, pp. 49–54. doi: 10.1145/1410012.1410024.

[4] F. Rivard, J. Bisson, F. Michaud, and D. Letourneau, "Ultrasonic relative positioning for multi-robot systems," in *2008 IEEE International Conference on Robotics and Automation*, May 2008, pp. 323–328. doi: 10.1109/ROBOT.2008.4543228.

[5] G. Benet, F. Blanes, J. E. Simó, and P. Pérez, "Using infrared sensors for distance measurement in mobile robots," *Robotics and Autonomous Systems*, vol. 40, no. 4, pp. 255–266, Sep. 2002, doi: 10.1016/S0921-8890(02)00271-3.

[6] M. T. Bui, R. Doskocil, and V. Krivanek, "The analysis of the effect of the parameters on indirect distance measurement using a digital camera," in *2019 International Conference on Military Technologies (ICMT)*, May 2019, pp. 1–9. doi: 10.1109/MILTECHS.2019.8870101.

[7] Z. Meng, X. Kong, L. Meng, and H. Tomiyama, "Stereo vision-based depth estimation," in *AIDE 2019: Advances in Artificial Intelligence and Data Engineering*, 2021, pp. 1209–1216. doi: 10.1007/978-981-15-3514-7_90.

[8] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig, "Monocular depth estimation using deep learning: A review," *Sensors*, vol. 22, no. 14, Jul. 2022, doi: 10.3390/s22145353.

[9] S. Liu, L. Zhao, and J. Li, "The applications and summary of three-dimensional reconstruction based on stereo vision," in *2012 International Conference on Industrial Control and Electronics Engineering*, Aug. 2012, pp. 620–623. doi: 10.1109/ICICEE.2012.168.

[10] A. Zaarane, I. Slimani, W. Al Okaishi, I. Atouf, and A. Hamdoun, "Distance measurement system for autonomous vehicles using stereo camera," *Array*, vol. 5, Mar. 2020, doi: 10.1016/j.array.2020.100016.

[11] Y. Pang, Y. Zhao, J. Chen, S. Wang, and H. Chen, "Viewing distance measurement using a single camera," in *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, Dec. 2014, pp. 512–515. doi: 10.1109/ITAIC.2014.7065103.

[12] T. Zhou, C. Sun, and S. Chen, "Monocular vision measurement system for the position and orientation of remote object," in *Proceedings of SPIE - The International Society for Optical Engineering*, Sep. 2007. doi: 10.1117/12.791428.

[13] J. V. G. Krishna, N. Manoharan, and B. S. Rani, "Estimation of distance to texture surface using complex log mapping," *Journal of Computer Applications*, vol. 3, no. 3, pp. 16–21, 2010.

[14] S.-H. Chang, C.-H. Hsia, W.-H. Chang, and J.-S. Chiang, "Self-localization based on monocular vision for humanoid robot," *Tamkang Journal of Science and Engineering*, vol. 14, no. 4, pp. 323–332, 2011.

[15] Y. Yang and Q.-X. Cao, "Monocular vision based 6D object localization for service robot's intelligent grasping," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1235–1241, Sep. 2012, doi: 10.1016/j.camwa.2012.03.067.

[16] Z. Zhang, Y. Han, Y. Zhou, and M. Dai, "A novel absolute localization estimation of a target with monocular vision," *Optik*, vol. 124, no. 12, pp. 1218–1223, Jun. 2013, doi: 10.1016/j.ijleo.2012.03.032.

[17] Muslikhin, D. Irmawati, F. Arifin, A. Nasuha, N. Hasanah, and Y. Indrihapsari, "Prediction of XYZ coordinates from an image using mono camera," *Journal of Physics: Conference Series*, vol. 1456, no. 1, Jan. 2020, doi: 10.1088/1742-6596/1456/1/012015.

[18] M. T. Bui, R. Doskocil, V. Krivanek, T. H. Ha, Y. T. Bergeon, and P. Kutilek, “Indirect method to estimate distance measurement based on single visual cameras,” in *2017 International Conference on Military Technologies (ICMT)*, May 2017, pp. 695–700. doi: 10.1109/MILTECHS.2017.7988846.

[19] Z. Zheng, X. Ren, and Y. Cheng, “An object distance measuring method of monocular vision mobile robot,” *International Journal of Control and Automation*, vol. 11, no. 3, pp. 117–128, Mar. 2018, doi: 10.14257/ijca.2018.11.3.11.

[20] J. Zhu and Y. Fang, “Learning object-specific distance from a monocular image,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 3838–3847. doi: 10.1109/ICCV.2019.00394.

[21] M. Vajgl, P. Hurtik, and T. Nejezchleba, “Dist-YOLO: Fast object detection with distance estimation,” *Applied Sciences*, vol. 12, no. 3, Jan. 2022, doi: 10.3390/app12031354.

[22] Pixy, “Pixy2 overview,” 2018. [Online]. Available: https://docs.pixycam.com/wiki/doku.php?id=wiki:v2:overview

[23] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics*. London: Springer London, 2009. doi: 10.1007/978-1-84628-642-1.

[24] H. Kim, C.-S. Lin, J. Song, and H. Chae, “Distance measurement using a single camera with a rotating mirror,” *International Journal of Control Automation and Systems*, vol. 3, no. 4, 2005.

[25] K. Molugaram, G. S. Rao, and A. C. Shah, “Curve fitting,” in *Statistical Techniques for Transportation Engineering*, Ken McCombs. Cambridge, MA, USA: Elsevier, 2017, pp. 281–288.

BIOGRAPHIES OF AUTHORS

Efemenah Endurance Idohen is a field operation manager at Globacom Limited, a telecommunication company in Nigeria. He holds a bachelor’s degree in mechanical engineering from the University of Benin and currently, he is pursuing his master’s degree in mechatronics at Ahmadu Bello University. Efemenah is a roboticist, and his research interests are in industrial/collaborative robotics, autonomous systems, human-machine interaction, and artificial intelligence. He can be contacted at efemenah@gmail.com.

Afolayan Matthew Olatunde is an associate professor at the Mechatronics/Mechanical Engineering Department of Ahmadu Bello University, Zaria, Nigeria. He holds a doctorate in robotics and a Master of Science degree in agricultural engineering from Ahmadu Bello University. His bachelor's degree is in Mechanical Engineering from Obafemi Awolowo University, Ile-Ife, Nigeria. His research interests are biomimetic robotics and mechatronic systems. He can be contacted at tunde_afolayan@yahoo.com or moafolayan@abu.edu.ng.

Umar Ali Umar is a senior lecturer in the Mechatronics Unit of the Mechanical Engineering Department at Ahmadu Bello University, Zaria. He has extensive research and industry experience in industrial automation, CAD/CAM, product design and development, and other advanced manufacturing technology and processes. He supervised several postgraduate students in Production Engineering and Mechatronics Engineering. He published several academic articles in reputable journals. He can be contacted at auumar@abu.edu.ng.