Artificial Life and Robotics (2023) 28:323–331
https://doi.org/10.1007/s10015-023-00856-0

ORIGINAL ARTICLE

Effect of display of YOLO's object recognition results to HMD for an operator controlling a mobile robot

Yuichi Sasaki[1] · Tetsushi Kamegawa[1] · Akio Gofuku[1]

Received: 11 April 2022 / Accepted: 22 January 2023 / Published online: 18 February 2023

Abstract
An operator feels a burden when he/she controls a rescue robot remotely because he/she has to keep watching camera images to find the target object. We think that this burden can be reduced by the combination of Head Mounted Display (HMD) and object recognition by deep learning. In the first half part of this study, we examine the effect that how presentation method by You Only Look Once (YOLO), a deep learning algorithm, and its recognition results to an operator wearing HMD. In the experiment, three methods of presentation were set: no display of object recognition, display only one object recognition result, and display 80 kinds of object recognition results. Under each presentation method, we measured the time it took for the operator to operate the robot and complete the given task. Additionally, we ask a questionnaire for each experiment. The results of the questionnaire showed that the method to present only one object recognition result was useful. In the second half part of this study, we develop a system to present 3D images with YOLO added, to further ease the burden of object search. Furthermore, we numerically prove that this system represents depth. In the experiment, two methods of displaying were set up: 2D images with Bounding Box (BB) by YOLO and 3D images with BB by YOLO. For each method of presentation, the operator operated the robot and recorded the number of objects found within a time limit. Additionally, we asked a questionnaire at the end of the search in each condition and at the end of all the experiments. The results of the questionnaire suggested points that need to be improved. Furthermore, we consider the flicker of the image found in the experiment.

Keywords Deep learning · Object recognition · HMD · 3D images

1 Introduction

When a large-scale disaster occurs, there is a need to quickly rescue people left on the scene. However, there are many disaster sites that are inaccessible to humans and where the risk of secondary disasters is high. In such cases, it is believed that understanding the environment inside the buildings in advance will make rescue operations safer and faster. Therefore, rescue robots are studied to explore the building safely and efficiently. This robot is controlled from a safe location by the operator and the robot explores the dangerous environment on behalf of the human. In Japan, many rescue robots have been developed since the Great Hanshin-Awaji Earthquake in 1995. Also, Quince, a disaster response robot, which was jointly developed by Tohoku University, Chiba Institute of Technology, and the International Rescue System, was actually deployed at the Fukushima Daiichi Nuclear Power Plant after the Great East Japan Earthquake that occurred on March 11, 2011 [1]. An operator feels a burden when he/she controls such robot remotely because he/she needs to keep watching camera image to find target objects.

A 3D stereoscopic using a stereo camera and Head Mounted Display (HMD) has been proposed as a remote operation method. This method has the advantage of preventing operator misrecognition in the Graphical User

This work was presented in part at the joint symposium of the 27th International Symposium on Artificial Life and Robotics, the 7th International Symposium on BioComplexity, and the 5th International Symposium on Swarm Behavior and Bio-Inspired Robotics (Online, January 25–27, 2022).

✉ Yuichi Sasaki
sasaki0y0mif@s.okayama-u.ac.jp

Tetsushi Kamegawa
kamegawa@okayama-u.ac.jp

Akio Gofuku
gofuku-a@okayama-u.ac.jp

[1] 3-1-1 Tsushimanaka, Kita-ku, Okayama-shi, Okayama 700-8530, Japan

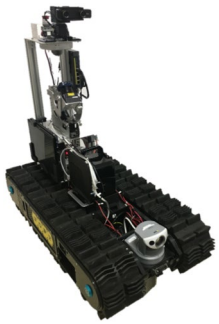

Fig. 1 Appearance of rescue robot DANIEL

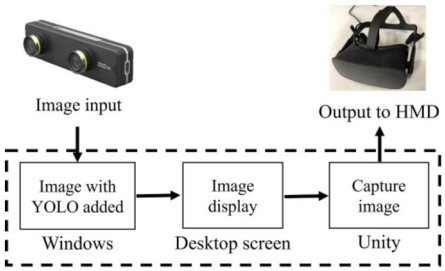

Fig. 3 System configuration of video image presentation

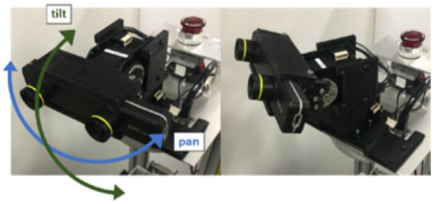

Fig. 2 Pan-tilt platform and stereo camera ZEDmini

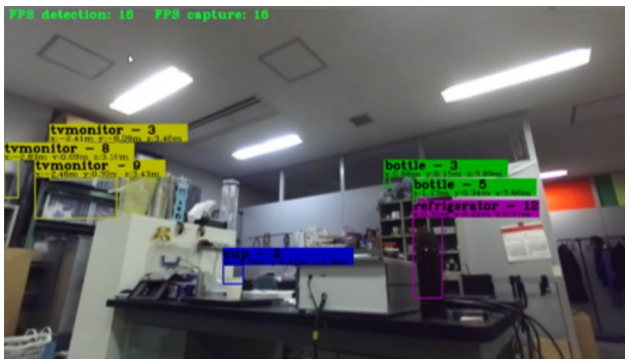

Fig. 4 Image seen by the operator

Interface (GUI) and its severe impact on the overall performance. Henrique et al. [2] compared the remote operation method using HMD and GUI. The results prove that the method using HMD is effective in object recognition. Additionally, recently, there has been various studies on the use of deep learning, which is expected to shoulder the burden of the operator [3]. In this study, we use You Only Look Once (YOLO), a deep learning object recognition algorithm invented by Joseph et al. [4].

This study aims to reduce the burden on the operator. In the first half of the paper, we examine the effect that how presentation method by YOLO, a deep learning algorithm, and its recognition results to an operator wearing HMD. In the second half of the paper, based on the previous result, we develop a system that displays 3D images with BB by YOLO, and we conduct an experiment to compare the 2D images with BB by YOLO.

2 Mobile robot remote control system

2.1 Mobile robot

In this study, we used DANIEL, a rescue robot owned by our laboratory. The appearance of DANIEL is shown in Fig. 1. A stereo camera (ZEDmini) and a pan-tilt head to mount it are mounted on the top of the aluminum frame behind DANIEL. The pan-tilt head is synchronized with the position of the HMD, as shown in Fig. 2. The pan-tilt head is constructed by combining two Dynamixel (MX-64R) heads, with a range of motion from − 110 to +110° for pan (horizontal direction) and −66 to +100° for tilt (vertical direction).

2.2 Image presentation system

In the first half of this study, the system presents 2D images to the operator using HMD. The version of YOLO employed is YOLOv4, which was developed by Alexey et al. [5], and we implement it on Windows operating system. The configuration of the image presentation system is shown in Fig. 3. When YOLOv4 is run on the input stereo camera images, the images after the run is displayed on the desktop screen. The images are captured in real time using Unity, a game development environment, and output to the HMD (Oculus Rift CV2). The image presented to the operator using the HMD is shown in Fig. 4.

3 Experiment

3.1 Experimental method

The experimental task was to remotely control DANIEL in the specified search area to find three objects and return to the starting position, and we measured the time between the start and the return. Each of the three objects to be

Fig. 5 Appearance of a paper cup

Table 1 Experimental conditions

α method	No display of object recognition
β method	Display only one object recognition result
γ method	Display 80 kinds of object recognition results

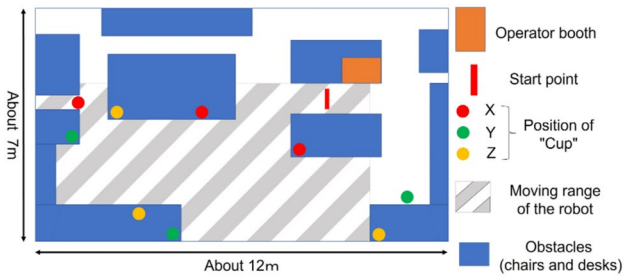

Fig. 6 Setting of experimental environment for the experiment

discovered is a paper cup (Cup) as shown in Fig. 5. In the experiment, the search area was specified, but the route to be searched by the operator was not particular. It should be noted that the robot is controlled by a game controller, and subjects are trained sufficiently in advance to be able to control the robot without directly seeing the game controller before conducting the experiment. At the end of each experiment, a questionnaire was administered. The questionnaires were evaluated on a 7-point scale and other comments in free text. The operators were 18 students at Okayama University. Each operator conducted the experiment three times. The operators wore HMD in all experiments, and they were presented different images for each experiment.

3.2 Experimental conditions

In the experiment, three presentation methods were set, as shown in Table 1. In β method, only Cup was trained using our own dataset, and in γ method, we used the training results on the Common Object in Context dataset.

3.3 Experimental environment

The experimental environment is shown in Fig. 6. According to the experimental conditions, three patterns of Cup placement were set. Three Cups were placed in each pattern. In all conditions, the placement positions were set to be recognized by the operator.

3.4 Questionnaire contents

The questionnaire contents are shown in the Table 2. In α method, three questions and free text were asked. In β and γ method, two questions about object recognition were added to the questions in α method.

3.5 Experimental results

Box plots and one-way analysis of variance at a significance level of 5% were conducted on the task completion time and on the scales obtained from Q1, Q2, Q3, and Q4 of the questionnaire. The experimental results are shown in Fig. 7. There was no significant difference in the task completion time. In the questionnaire, significant differences were recorded in Q2, Q3, and Q4. Therefore, multiple comparisons using the Tukey method were conducted for Q2 and Q3. This multiple comparison showed significant differences between methods α and β and methods β and γ in Q2, and between methods α and γ and methods β and γ in Q3. In Q5, 44.44% and 27.78% of the total respondents answered yes for method β and γ, respectively.

Table 2 Questionnaire contents

α	β, γ	Q1	Did you feel comfortable operating the mobile robot?
		Q2	Was it easy to search for the object?
		Q3	Was it easy to recognize not only the object but also the surrounding environment?
		Q4	Did you feel that the Bounding Boxes interfered with the operation?
		Q5	Was it faster for you to recognize an object on the screen than to find it by yourself?
Other comments(about 3D sickness, good and bad points of the system, etc.)			

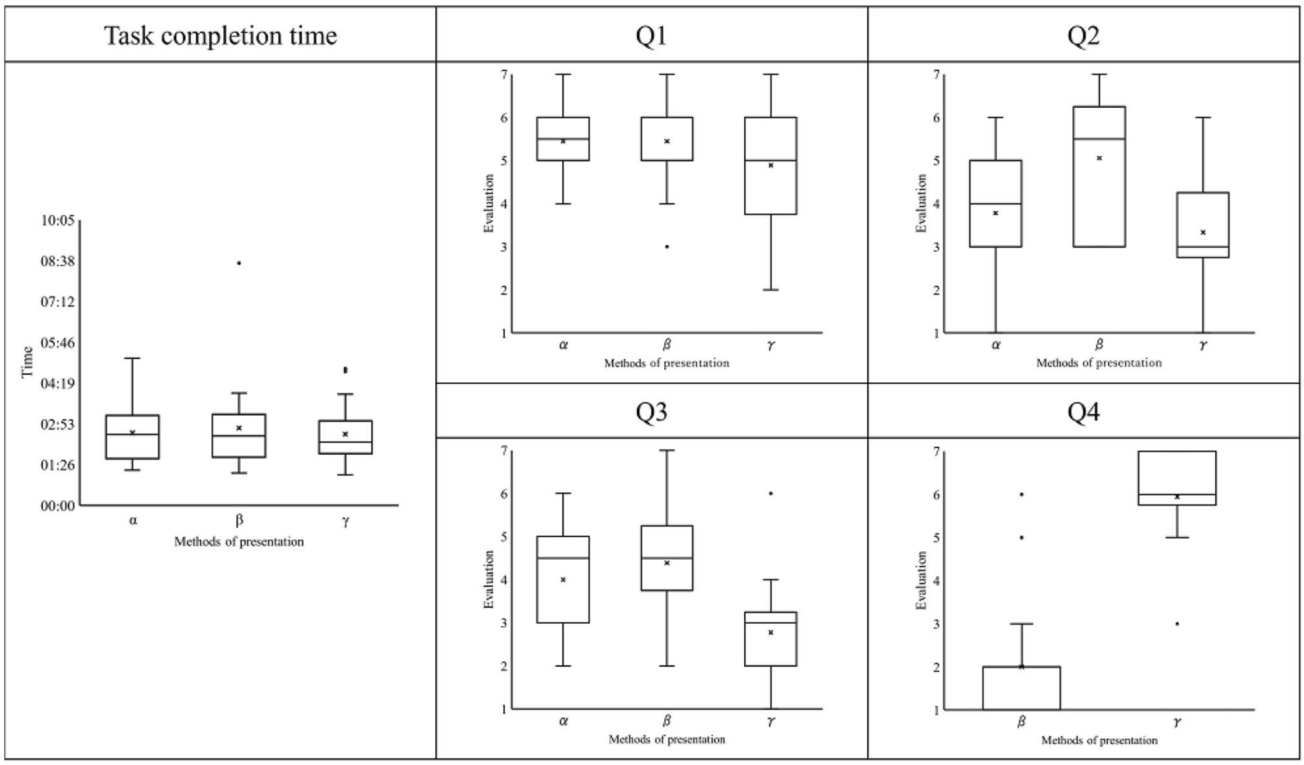

Fig. 7 Box plots for task completion time and questionnaire Q1, Q2, Q3, Q4

3.6 Consideration

We think that the reason why there is no significant difference in task completion time is that because the task is easy. The experimental task was to find three Cups and return to the starting point. The average completion time was 2 min and 36 s. Therefore, we think that no difference was generated because most of the task completion time was operation time.

The results of the questionnaires were considered. The results of the analysis for each question suggest that method β has a positive influence on the operator in object recognition. We consider the details of each question below.

- Q1: Did you feel comfortable operating the mobile robot?
  This question was asked to ensure that the operation would not be affected by the methods. From the results, it was confirmed that the operation was not affected by the methods. However, many people indicated that the box plot was difficult to operate under method γ. We think the reason is because in our experiment, many people operated while looking at the crawler when moving, and YOLO's recognition responded to the crawler. The situation is shown in Fig. 8.
- Q2: Was it easy to search for the object?

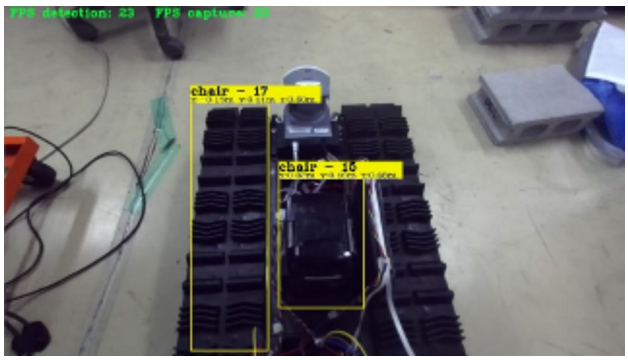

Fig. 8 Recognition of the crawler part under condition γ

The results showed that there was a significant difference between methods α and β, and between methods β and γ. Therefore, it is clear that the operators felt it was easier to search for objects in method β than in another method. From the free text, in method β, the respondents answered that "There were times when the label became an assistance" and "I could take the steps to check myself after the label responded." This experiment aims to examine a presentation method to support the operator's concentration, which was reduced by long hours of remote operation, so the expected results were obtained. Then, we compare methods β and γ. In method γ, the

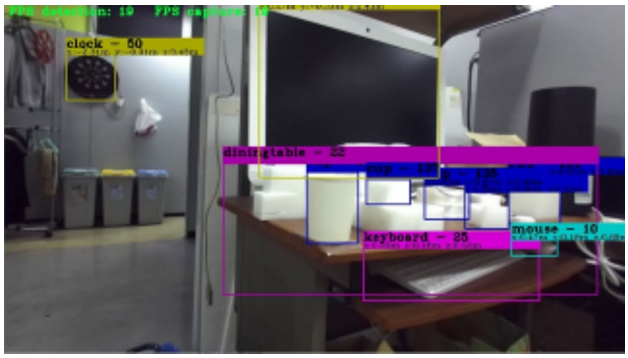

Fig. 9 Example of over recognition under condition γ

respondent answered "I didn't notice the recognition of the Cup." We think that this response is because numerous labels of object recognition results were displayed in a crowded location. Thus, the Cup was hidden. The situation is shown in Fig. 9. Therefore, method β, display only one object recognition result that is considered effective.

- Q3: Was it easy to recognize not only the object but also the surrounding environment?

  The results showed that there was a significant difference between methods α and γ, and between methods β and γ. Therefore, the operators felt it was difficult to recognize the surrounding environment in method γ. We compare methods α and γ. From the free text, in method γ, many respondents answered "The environment was difficult to see because there were too many labels." This reason for such response could be because the labels interfered with the recognition of the environment. This is the same as in Q2. Then, we compare methods β and γ. In method β, there were few operators who felt it was difficult to recognize the environment as in method α, because they would not recognize the object unless the camera captured it. However, one respondent answered, "When the system misrecognizes, it is attracted to the label, because the label is of a kind." We think that this is response is because operator's concentration was focused on the labels that suddenly appeared on the display.
- Q4: Did you feel that the Bounding Boxes interfered with the operation?

  The results showed that there was a significant difference between methods α and γ. Therefore, it is clear the operators felt that the number of people who felt that the label was an obstacle in method β was small. We think the reason for this outcome is that method γ displays 80 kinds of labels, while method β displays only one kind. Because of the experimental environment, method γ keep displaying some recognition result in most cases, and we think it interfered with the operation.
- Q5: Was it faster for you to recognize an object on the screen than to find it by yourself?

  The results showed that method α requires the operator to find the object by him/herself, while methods β and γ, which use deep learning object recognition, support the operator's search. However, the difference between methods β and γ indicates that the effect would be reduced if the presentation method to the operator is inappropriate.
- Other comments(about 3D sickness, good and bad points of the system, etc.)

  In the free text, there were many respondents who said that they got sickness. It is proven that long-term immersion using HMD can worsen motion sickness. However, our experiment did not take a long time to conduct. The common factor among the methods is that the presentation method was 2D screen. This presentation method differs from reality, where the depth can be felt, and depth cannot be grasped. Therefore, I think this uncomfortable feeling is what caused the sickness.

4 3D images with added YOLO's object recognition results

From the experiment in Sect. 3, I think that the cause of sickness was the 2D images. Therefore, we developed a system to present 3D images that added YOLO's object recognition results. The system uses a stereo camera (ZEDmini) and a HMD (Oculus Quest2). The configuration of the system is shown in Fig. 10. The image that the operator sees by the running result is shown in Figs. 11 and 12.

The following sections describe the details.

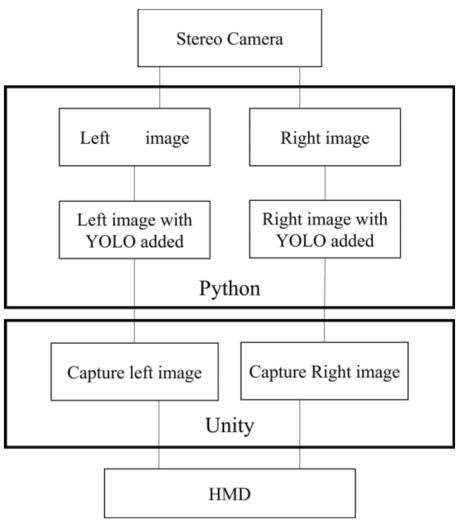

Fig. 10 System configuration

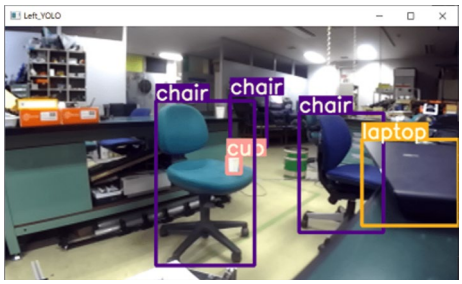

Fig. 11 Image seen by the operator with the left eye

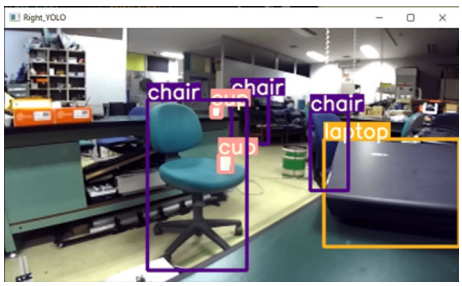

Fig. 12 Image seen by the operator with the right eye

4.1 Trimming the images

To provide stereoscopic images using a stereo camera and HMD, it is necessary to output the left and right side image acquired by the stereo camera to the left and right eye of the HMD respectively. However, the images acquired by ZED-mini are in a side-by-side format. Therefore, it was necessary to trim the image in the center and divide it into two images, which can be done with a Python program.

4.2 Add YOLOv3

Here, we add YOLO to each of the two trimmed images. The same algorithm is used for the left and right images, but the programs themselves are independent. The system uses YOLOv3 in honor of Joseph Redmon.

4.3 Output to HMD

Unity is used to output the video to the HMD. For this purpose, the left and right images with YOLO added are each acquired as game objects in it. It is necessary that the left and right images are displayed in the center of the HMD screen whenever the operator wearing the HMD faces any direction. Therefore, we constructed such a scene in Unity. The structure of the scene is shown in Fig. 13. For example, let us take "Left." The left eye of the HMD is placed where it can capture the image of the left camera. The solid lines represent objects that are displayed on the game screen, and

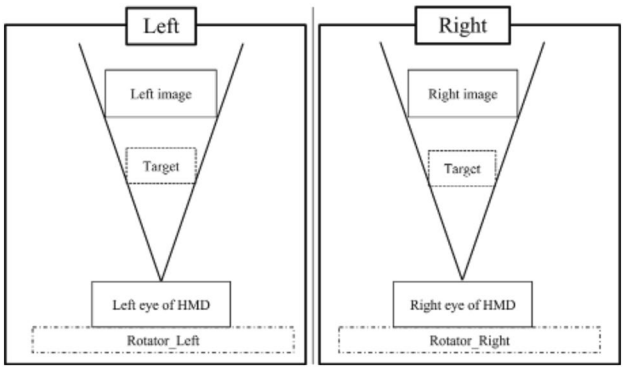

Fig. 13 System configuration of Unity

the dotted lines represent objects that are not displayed. The left camera image is a child object of Rotator_Left, so the left camera image always keeps a constant distance from Rotator_Left. Rotator_Left is attached with a script that keeps acquiring the rotation of the left eye of the HMD, so the rotation of the left eye of the HMD and the rotation of Rotator_Left are synchronized. The target object is provided to refer to the initial position. These are the same for Right. The left and right images are divided into layers to limit the drawing range. Therefore, the left camera images are always displayed on the left and right eye of the HMD, respectively.

5 Verification of the constructed system

In this chapter, we conduct verification using measured disparity and preliminary experiment in which images are presented to subjects to confirm that images are converted to 3D in the constructed system.

5.1 Disparity

Disparity is the difference in the position of the corresponding area between two images. The smaller the disparity, the farther back the object is, and the larger the disparity, the farther forward the object is. This is proven by the principle of triangulation.

5.2 Disparity calculation

The center-of-gravity coordinates of the BB were used to calculate the disparity. Therefore, the parameter for calculating the disparity is the x-coordinate of the center-of-gravity coordinates. Examples of images and numerical values used to calculate the disparity are shown in Fig. 14 and Table 3. When the center-of-gravity coordinates of the BB are used to calculate the disparity, the size of the left and right BBs must be the same degree. The size of

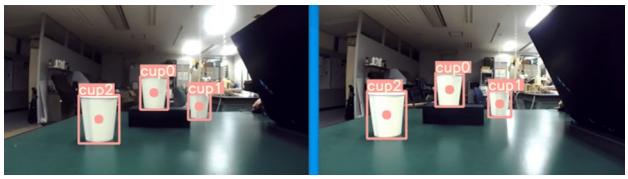

Fig. 14 Image used for disparity calculation

Table 3 Center-of-gravity coordinates

		center_x	center_y
Left	Cup0	332.5	192.5
	Cup1	431.5	222.5
	Cup2	211.5	250.0
Right	Cup0	286.0	187.0
	Cup1	396.0	218.0
	Cup2	148.0	242.0

Table 4 Size of bounding box

		sidelength_x	sidelength_y
Left	Cup0	61.0	73.0
	Cup1	49.0	57.0
	Cup2	89.0	106.0
Right	Cup0	62.0	74.0
	Cup1	50.0	58.0
	Cup2	86.0	104.0

Table 5 Disparity calculation

	Cup0	Cup1	Cup2
Disparity	46.5	35.5	63.5

the BB in Fig. 14 is shown in Table 4. The image size was 672×376 (px). From Table 4, the difference in the size of one side in the left and right BBs was at most 3 px, and there is no big difference between the left and right BBs. So, there is no problem in using the center-of-gravity coordinates to calculate the disparity. It should be noted that as a preliminary experiment to examine the effectiveness of the disparity measurement based on YOLO's BB, we conducted an experiment using a laptop PC as the target object in addition to a paper cup and confirmed that the shape of the detected object has almost no effect on the measured disparity. Based on the above, the disparity calculated from Table 3 is shown in Table 5. The results show that Cup1, Cup0, and Cup2 are present from the back. Moreover, in reality, the paper cups are set up in such a way. Therefore, we could prove that the constructed system represents the depth.

5.3 Preliminary experiment

Several subjects were asked to use the constructed system. As a result, the images were made in 3D, and the subjects could feel the depth of the BBs.

6 Additional experiment

In this chapter, we present an experiment that examined whether there is a difference between displaying 3D images and 2D images with presentation of BBs by YOLO.

6.1 Experimental method

The task for the subjects is to remotely control the mobile robot DANIEL for two minutes to search for an object within a designated search area. The object is a paper cup, and the number of paper cups found is recorded. It should be noted that although six paper cups are placed in the environment, the subjects are not told in advance how many paper cups are placed in the experimental environment. Incidentally, there was no case in which the subject was able to search all the paper cups within the time limit. In the experiment, the search area is specified, but the search route is free. The subjects are 12 students in our laboratory, and each subject is asked to explore the environment by using 3D and 2D images one at a time. Here, images presented to the subjects are added BBs with YOLO in both conditions. The operation of the robot is the same as in the experiment in Chapter 3, and the subjects perform the experiment after prior training. Subjects are also asked to complete a questionnaire at the end of each search under certain conditions and at the end of all experiments. The questionnaire asked to the subjects at the end of each search is that to answer questions on a scale of 1 to 7. The questionnaire asked to the subjects at the end of all experiments is that to select whether they preferred 2D or 3D, or whether there was no difference between the two, and to answer the questions with free comments. The Table 6 shows the specific questionnaire.

6.2 Experimental conditions and environment

To eliminate the influence of habituation, subjects were divided into two groups: one group of 6 subjects who explored first by 3D images and the other group of 6 subjects who explored by 2D images first. In addition, two different Cup positions were set to avoid the same position in the two searches. Accordingly, we divided the group into two groups, one group of 6 subjects who searched for position pattern 1 first and the other group of 6 subjects who searched for position pattern 2 first, to consider the effect of the difference in difficulty level due to the Cup

Table 6 Questionnaire contents

At the end of each search	Q1	Did you feel comfortable operating the mobile robot?
	Q2	Was it easy to search for the object?
	Q3	Was it easy to recognize not only the object but also the surrounding environment?
	Q4	Did you feel sickness?
At the end of all experiments	Q1	Which presentation method made it easier to search for the object?
	Q2	Which presentation method made it easier for you to recognize the environment, not only the object?
	Q3	Which presentation method enabled you to operate with an understanding of distance (depth)?
	Other comments (good and bad points of the system, etc.)	

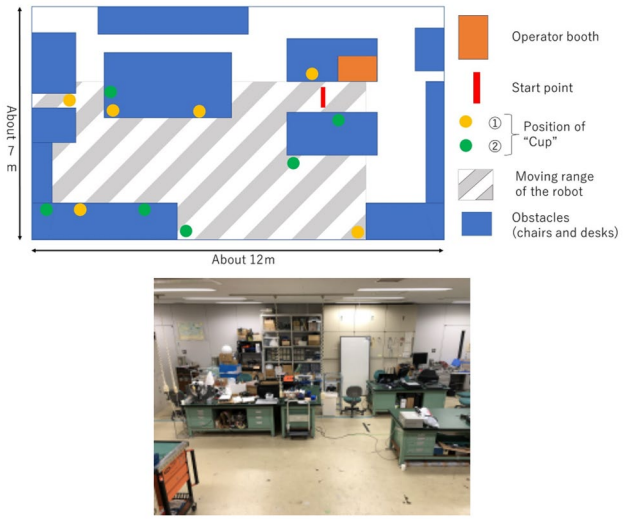

Fig. 15 (Upper) Setting of experimental environment for the additional experiment (Lower) Picture of the experimental environment

Table 7 Totalization of the results of the questionnaire at the end of all experiments

Q1	3D is better	6
	2D is better	5
	No difference	1
Q2	3D is better	7
	2D is better	4
	No difference	1
Q3	3D is better	10
	2D is better	1
	No difference	1

placement. In other words, there are four experimental conditions in total, with 3 subjects assigned to each experimental condition. The specific experimental environment and the positions of the Cups are shown in Fig. 15. Cup positions are those that subjects can explore without any problems by looking at the camera images of the mobile robot.

Table 8 Recognition pattern

Object recognition results		Left and right recognition patterns	
Left	Right	AND Pattern	OR Pattern
OK	OK	OK	OK
OK	NO	NO	OK
NO	OK	NO	OK
NO	NO	NO	NO

6.3 Experimental results

First, a t-test was conducted at the 5% significance level on the number of Cups found by the 2D and 3D search and on the questionnaire contents at the end of the search, and no significant differences were found. Next, the Table 7 shows the totalization of the results of the questionnaire at the end of all experiments.

From the free comments, we obtained the following comment: "It was easier to find objects in 3D, but 2D was easier to operate because it was harder to get sickness." The results of the questionnaire suggested the effectiveness of the 3D images as well as points that need to be improved. Furthermore, this experiment shows that in 3D images with YOLO added, the results recognized by the left and right cameras are not always displayed at the same timing, and in this case, the operator feels flickering in the image.

6.4 Consideration

This experimental examination did not necessarily indicate that 3D is more useful. A report surveying over 150 previous studies on robot teleoperation [6] states that stereo cameras and stereoscopic displays are particularly effective in confined spaces such as the collapsed World Trade Center [7]. Therefore, in a relatively wide and monotonous search environment such as the one set up in this experiment, we think that there was no significant difference between 2D and 3D.

Next, we examine how to prevent images from flickering when displayed in 3D. The Table 8 shows the patterns that

can be presented to the operator when the recognition results differ between the left and right images. To simply prevent flickering of the image, a method of outputting it to the operator only when the same recognition result is output on the left and right sides could be considered. This recognition pattern is called the AND Pattern. However, if the images presented to the operator are only AND Pattern, the operator will not experience flickering, but only results that match the recognition results of both eyes will be output, resulting in fewer objects being output and the possibility of missing the object to be searched. On the other hand, the method used in this experiment, which presents the results to the operator even when the recognition results differ between the left and right images, outputs all object recognition results in the left and right images, even though the operator may experience flickering, so the possibility that the operator will miss the target object is low. This recognition pattern is called the OR Pattern. Considering the merits and demerits of the AND Pattern and OR Pattern in consideration of the operator's recognition ability, the AND Pattern is effective when the number of recognition targets is large, while the OR Pattern is effective when the number of recognition targets is small. Therefore, it is appropriate to set a threshold value according to the number of recognition target objects in the environment and switch between AND Pattern and OR Pattern to present BBs to an operator.

7 Conclusion

In the first half part of this paper, we examined the effect of presentation method by YOLO, a deep learning algorithm, and its recognition results to an operator wearing HMD. Three methods of presentation were set for the experiment: no display of object recognition as method α, display only one object recognition result as method β, and display 80 kinds of object recognition results as method γ. The step in the experiment are as follows: (1) remotely controlling a mobile robot in a specified search area, (2) finding three Cups, and (3) returning to the starting position. The questionnaire results indicated that the method β, which displays only one object recognition result, was useful in this experiment. In the second half of this paper, we constructed a system to present 3D images with YOLO added. We explained the structure of the system and numerically proved that the depth of the system was represented. In the experiment, two methods of displaying were set up: 2D images with BBs by YOLO and 3D images with BBs by YOLO. The steps of the experiment consisted of remotely operating a mobile robot within a specified search area and finding as many cups as possible within a time limit.

Funding Open access funding provided by Okayama University.

Data Availability There is no data to share to readers in this paper.

References

1. Koyanagi Eiji, Tomoaki Yoshida, Nishimura Takeshi (2013) Robot system for indoor exploration of Fukushima Daiichi nuclear power plant. J Robot Soc Jpn 31(1):47–48 (in Japanese)
2. Henrique M, Rodrigo V (2009) Immersive 3-D teleoperation of a search and rescue robot using a head-mounted display. In: IEEE conference on emerging technologies & factory automation
3. Matsumura Yuto, Kamegawa Tetsushi, Gofuku Akio (2018) Prototype of Object Recognition System by Deep Learning using Camera Image Acquired by Mobile Robot. In: 19th SICE System Integration Division Annual Conference. (in Japanese)
4. Joseph R, Santosh D, Ross G, Ali F (2016) You only look once: unified, real-time object detection. Cornell University arXiv:1506.02604
5. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: optimal speed and accuracy of object detection. Cornell University arXiv:2004.10934
6. Chen Jessie Y. C, Haas Ellen C, Barnes Michael J (2007) Human performance issues and user interface design for teleoperated robots. IEEE Trans Syst Man Cybern Part C Appl Rev 37(6):1231–1245
7. Murphy Robin Roberson (2004) Human-robot interaction in rescue robotics. IEEE Trans Syst Man Cybern Part C Appl Rev 34(2):138–153