A Method of Improving the QOL of the People with Visual Impairment by MY VISION

Shun Kitazumi
Graduate School of Engineering, Kyushu Institute of Technology, 1-1 Sensuicho, Tobata-ku, Kitakyushu, 804-8550, Japan

Yui Tanjo
Faculty of Engineering of Engineering, Kyushu Institute of Technology, 1-1 Sensuicho, Tobata-ku, Kitakyushu, 804-8550, Japan
Email: kitazumi.shun215@mail.kyutech.jp, tanjo@cntl.kyutech.ac.jp

Abstract

Visually impaired people face several difficulties in indoor activities, such as spending excessive time in locating objects. This paper proposes a method for assisting object acquisition by detecting desired objects and guiding users to them. The method requests a user to express the object he/she wants to acquire verbally and utilizes speech recognition to detect the specified object. Subsequently, the system guides in voice the user's hand to the location of the desired object. The performance of the method is experimentally shown. The method contributes to enhancing the comfort of indoor activities of visually impaired and, in this way, improves their quality of life.

Keywords: MY VISION, Visually impaired, QOL, Object acquisition, Network diagram

1. Introduction

Many people around the world suffer from visually impaired vision, a condition that interferes with their daily lives due to persistent vision loss. According to a report [1] published by the World Health Organization (WHO) in 2019, the number of people with visual difficulties worldwide is estimated to be at least 2.2 billion, and there are concerns that the number of people affected will increase further due to population growth and aging.

In response to this worldwide social problem of visibility difficulties, a great deal of research and development has been conducted to support the daily lives of people with visibility difficulties. For example, there has been the development of a pedestrian crossing navigation system using smart glasses to enable people with vision difficulties to safely walk across pedestrian crossings alone [2][3][4][5], and the development of a system to assist people in getting on and off public transportation [6][7].

As in the above studies, much of the research on assisting people with vision difficulties tends to focus on supporting outdoor activities. However, there are many problems that occur during indoor activities, such as spending excessive time for acquiring an object when looking for it at home. Therefore, the purpose of this study is to improve the quality of life (QOL) of people with visual acuity difficulties by focusing on a proposal for an object acquisition support method, which plays a part in supporting indoor activities for people with visual acuity difficulties.

Currently, an application has been released that uses a smartphone to capture a registered object with a camera and read it aloud to the user[8]. This application is very useful for identifying objects that are similar in shape, but since it assumes that the object's location can be accurately identified and captured by the camera, it does not work unless the visually impaired person knows the object's location. In addition, as a study that addresses the guidance of object acquisition paths, a system has been proposed by J.P. Docto *et al.* [9], in which a smart glove is worn by a visually impaired person and guides him or her to the object utilizing the built-in camera and sensors. This system does not require text input or touch operation and is easy for visually impaired people to handle. However, in practical use, there are issues related to practicality, such as the time and effort required to put on and take off the gloves, the weight of the gloves when they are worn, and the resistance caused by the wiring.

In this study, we propose a system that provides voice guidance on a route to the object requested based on MY VISION (a Magic eYe of a Visually Impaired for Safety and Independent actiON), thereby reducing the burden of visually impaired people and making the system more intuitive and practical. Specifically, the system can provide real-time guidance using a notebook PC and a webcam, regardless of location, and enables it by voice without text input or button operation. This is expected to solve the practical limitations of smartphone applications and smart gloves that have been pointed out in previous research. Therefore, the system proposed in this study can be an important step toward improving the quality of life of people with vision difficulties during indoor activities by supporting smooth acquisition of the target object when they are looking for something at home.

2. Methodology

2.1. Voice recognition

In this study, when a user has an object that he/she wants to acquire, the user is asked to tell the object loud (speech transmission) to the proposed system, and the target object's name is recognized by analyzing the speech. By asking the user to say the trigger word before the object name, the noun that follows the trigger word is known as the target object's name. In the system, "MY VISION" is used as the trigger word. For example, if the user wants to acquire a cell phone, he/she just say, "MY VISION, find a cell phone". The system then guides the user's hand to the location of the desired object using voice guidance.

2.2. *Object detection*

In the proposed system, we use YOLOv7 (You Only Look Once version 7) [10] to detect a target object in real time. Note that in the system, objects that are present on the desk and easily portable (e.g., remote controls, cell phones, cups) are considered as a class of objects to be detected. In addition, by combining YOLOv7 and DeepSORT (Deep Learning Simple Online and Realtime Tracking) [11], object features are extracted on each frame, and each object is assigned a unique ID for tracking. Fig. 1 shows the object detection results.

2.3. *Route guidance to a target object*

If a target object is detected, the direction of the object location (referred to as a route hereafter) is guided by voice guidance according to the clock position. We use MediaPipe Hands [12] to estimate the user's hand area and align the user's hand with the target object. When the target object is in the 9 o'clock direction (left direction), the system announces "move to 9 o'clock direction" with voice.

The proposed method not only focuses on the hand and the target object, but also employs proposed a network diagram to represent the positional relationship among the surrounding objects based on the object detection results. In the network diagram, nodes are defined as the center coordinates of the object, and the edges are the distances in pixel between objects. The network diagram is introduced because the system can find the target object position smoothly, and also can find the position of the target object more precisely, when the target is very close to the other objects.

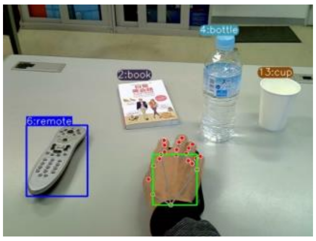


Fig. 1 Object detection results. (The target object is drawn in a blue bounding box and the user's hand is in a green box.)

Moreover, if an object to be detected is not detected in the current frame, then the system estimates the undetected position by referring to the previous network diagram.

The first condition for selecting the network diagram to be referred to is when the number of detected objects is three or more and the number of detected objects is the largest (Eq. (1)), and the second condition is when the network diagram consists of the same type of objects for a certain period (Eq. (2)).

$$\begin{cases} obj\_cnt_{F(t)} \geq 3 \\ obj\_cnt_{F(t)} \geq max\_obj\_cnt \end{cases} \tag{1}$$

Here, $F(t)$, $obj\_cnt_{F(t)}$ and *max_object_cnt* are current frame, the number of objects detected in the current frame and the maximum number of objects detected in each frame, respectively.

$$S = \begin{cases} S+1 & \text{if } \forall E\big(F(t)\big) = \forall E(F(t-1)) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$S \geq Th \Longrightarrow \text{SetReferenceModel}(\forall E(F(t)))$$

Here, S, $E(F(t))$, and Th are the variable for state continuity, node in $F(t)$ and threshold, respectively.

The object location estimation method first calculates the amount of object movement common to the current frame and the reference network diagram, respectively, and derives the average amount of movement (Eq. (3)). The position of the undetected object is then estimated based on the coordinates of the model and the average displacement (Eq. (4)). The meanings of the letters in the equations are shown in Table 2.

$$\begin{cases} \bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(x_{t,i} - x_{model,i}\right) \\ \bar{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(y_{t,i} - y_{model,i}\right) \end{cases} \tag{3}$$
$$(i = 1, \cdots, n)$$

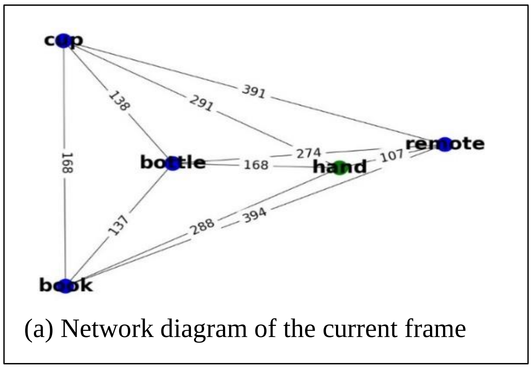
$$\begin{cases} x_{est,k} = x_{model,k} + \bar{x} \\ y_{est,k} = y_{model,k} + \bar{y} \end{cases} \tag{4}$$

Table 2. The meanings of Eq. (3) and Eq. (4)

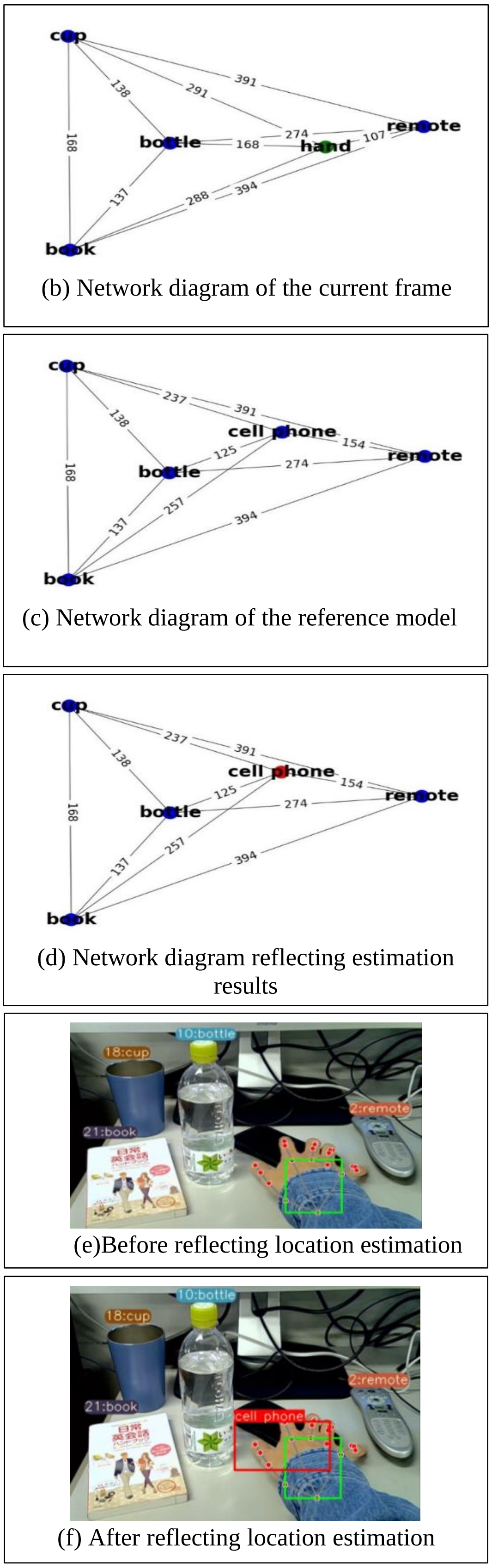
$\bar{x}, \bar{y}$	Average displacement in the x or y-axis direction
n	Number of common nodes between current frame and reference model
$x_{t,i}, y_{t,i}$	x or y coordinates of the common node in the current frame
$x_{model,i}, y_{model,i}$	x or y coordinate of the common node in the model
$x_{est,k}, y_{est,k}$	x or y coordinate estimates for missing nodes
$x_{model,k}, y_{model,k}$	x or y coordinates of the node in the model corresponding to $x_{est,k}$ or $y_{est,k}$

The first condition for starting object location estimation is when a reference network diagram exists, and the second condition is when the number of detected objects is two or more and there are two or more objects in common with the reference network diagram. Fig. 2 shows the results of object location estimation using the above condition.

In the proposed system, a monocular RGB camera is used to align the hand with the target object without using a depth camera. The condition for determining object acquisition is achieved by employing the bounding box positional relationship between the target object and the hand region, and by judging the change in hand orientation that occurs when the hand gestures come closer to the camera after grasping the object. By adding the latter condition, it is expected to prevent the error judgment of acquisition as successful simply because the bounding boxes of the two objects overlap, even when the depths are different, and the objects are not touched. Furthermore, the system also incorporates a process that checks whether or not the object closest to the hand matches the target object by referring to the network diagram and announces when the subject has acquired the wrong object. When the above three conditions are satisfied, the target object is judged to have been acquired, and the acquisition is communicated by voice as completed. Fig. 3 shows an example that the object acquisition decision was fulfilled.



(a) Network diagram of the current frame



(b) Network diagram of the current frame

(c) Network diagram of the reference model

(d) Network diagram reflecting estimation results

(e)Before reflecting location estimation

(f) After reflecting location estimation

Fig. 2 Result of object location estimation
(When the target object is estimated, it is surrounded by a red bounding box.)

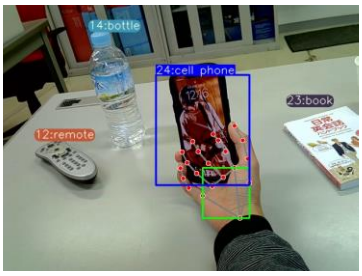


Fig. 3 Image on completion of the object acquisition that satisfies the three conditions

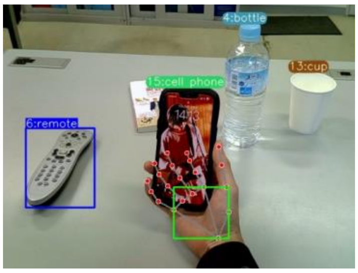


Fig. 4 Arrangement where other objects are present in the guidance path to the target object (target object: remote controller)

3. Experimental Result

Five objects were selected as candidate objects for acquisition: a book, a cell phone, a cup, a plastic bottle, and a remote controller. An experiment was conducted to select an acquisition target from these objects and to make a robot guide a blindfolded user along a path to the target object. As shown in Fig. 4, we placed other objects on the guided path to the target object to verify if it is possible to approach the target object again after announcing that the wrong object has been grabbed. The target object position was changed in the directions of 9, 10, 12, 2, and 3 o'clock with respect to the hand position to ensure that there was no dependence on the object placement. The evaluation index was *Accuracy* which represents the percentage of correct answers acquired. A total of 25 acquisition experiments were conducted with 5 choices among 5 objects times 5 object positions, resulting in an Accuracy of 92%.

4. Conclusion

In this paper, we proposed a system that provides voice guidance on the path to object acquisition based on MY VISION and contributes to improving the quality of life during indoor activities of the people with vision difficulties. In addition, by utilizing the information obtained from Mediapipe Hands and network diagrams, we developed a practical and intuitive system that provides all information and guidance by voice, as well as countermeasures in case that object tracking is interrupted.

References

1. WHO: “"World report on vision", https://www.who.int/publications/i/item/9789241516570, 2019, (Accessed 2023-12-04).
2. H. Son, D. Krishnagiri, V. Jeganathan, J. Weiland: “Crosswalk guidance system for the blind”, 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp.3327-3330, 2020.
3. J. K. Tan, T. Ishimine, S. Arimasu: “Walk Environment Analysis Using MY VISION: Toward a Navigation System Providing Visual Assistance”, International Journal of Innovative Computing, Information and Control, Vol. 15, No.3, pp.861-871, 2019.
4. J. K. Tan, K. Kitagawa: “A Method of Navigating a Visually Impaired Person Using MY VISION”, Journal of Robotics, Networking and Artificial Life, Vol.9, No. 1, pp. 25-30, 2022.
5. T. Kumano, J. K. Tan: “Traffic signs and signals detection employing the my vision system for a visually impaired person”, ICIC Express Letters, Part B: Applications pp. 385-391, 2015.
6. S. M. Cruz, L. A. M. Hernández, et al.: “An Outdoor Navigation Assistance System for Visually Impaired People in Public Transportation", IEEE Access, vol.9, pp.130767-130777, 2021.
7. J. K. Tan, Y. Hamasaki, Y. Zhou, I. Kazuma: “A Method of Identifying a Public Bus Route Number Employing MY VISION”, Journal of Robotics, Networking and Artificial Life Vol.8, No.3, pp.224-228, 2021.
8. INTEC Inc. : https://www.intec.co.jp/news/2020/0318_1.html, (Accessed 2023-12-05).
9. J. P. Docto, A. I. Labininay, J. F. Villaverde: “Third eye hand glove object detection for visually impaired using You Only Look Once (YOLO)v4-tiny algorithm”, 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pp.1-6, 2022.
10. Y. Wang, A. Bochkovskiy, H. Liao: “"YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors”, arXiv preprint arXiv:2207.02696, 2022.
11. N. Wojke, A. Bewley, D. Paulus, “Simple online and realtime tracking with a deep association metric”, 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649, 2017.
12. MediaPipe, https://developers.google.com/mediapipe, (Accessed 2023-12-05).

Authors Introduction

Mr. Shun Kitazumi

He received his Bachelor’s degree in Engineering in 2022 from the Faculty of Engineering, Kyushu Institute of technology in Japan. He is currently a master student in Kyushu Institute of Technology, Japan

Dr. Yui Tanjo

Dr. Tanjo is currently a professor with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology. Her current research interests include ego-motion analysis by MY VISION, three-dimensional shape/motion recovery, human detection, and its motion analysis from video. She was awarded SICE Kyushu Branch Young Author’s Award in 1999, the AROB Young Author’s Award in 2004, the Young Author’s Award from IPSJ of Kyushu Branch in 2004, and the BMFSA Best Paper Award in 2008, 2010, 2013 and 2015. She is a member of IEEE, The Information Processing Society, The Institute of Electronics, Information and Communication Engineers of Japan.