Sensing and Navigation of Wearable Assistance Cognitive Systems for the Visually Impaired

Guoxin Li, Jiaqi Xu, Zhijun Li, *Fellow, IEEE*, Chao Chen, and Zhen Kan, *Member, IEEE*

***Abstract*—This article develops a wearable vision-based assistance system to provide situational awareness for blind and visually impaired (BVI) people in indoor scenarios. The system is built upon nonintrusive wearable devices, including an RGB-D camera, an embedded computer, and haptic modules. First, the depth map and color images of the scene are obtained from an RGB-D camera, which provides 3-D environmental information. The modular work modes are then designed for different tasks, such as navigation and multitarget recognition. Then, the cognition results are summarized and presented to the user through verbal or haptic feedback. Our system is evaluated by a pilot test to validate its effectiveness of improving the navigation capabilities and multitarget recognition capabilities for the BVI in indoor environments. We present study results with different tasks, including navigation, object localization, face recognition, and text reading. The experiments prove that the system can meet the needs of the BVI in daily use.**

***Index Terms*—Human–computer interaction, multitarget recognition, simultaneous localization and mapping (SLAM).**

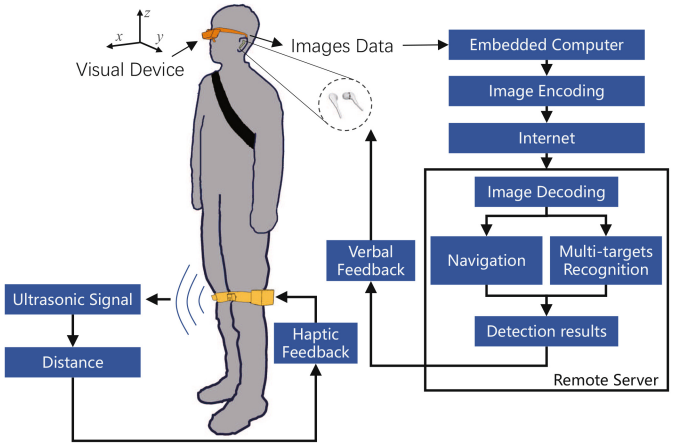


Fig. 1. Overview of the wearable assistance system.

I. Introduction

VISUAL impairment is a long-standing problem that affects millions of people worldwide. According to the World Health Organization, 285 million people are living with visual impairment, which makes the blind and visually impaired (BVI) difficult to travel freely [1]. Canes or guide dogs can be helpful in avoiding obstacles when the BVI walk. However, canes fail to proactively provide rich information feedback about the surrounding environment, which causes the BVI reduced mobility and increased possibility of accidental injury such as falls. A well-trained guide dog can offer assisted walking by steering the BVI around obstacles and alerting them to danger by barking in unfamiliar routes. But the cost of training a guide dog can be expensive and thus, prevents the availability. Therefore, advanced visual assistance devices that can provide the BVI with environment perception and navigation capability are desired.

With recent advances in artificial intelligence and wearable devices, many technological solutions have been developed and used to help the BVI. In recent years, there has been an increasing demand to improve the individual's quality of life using assistive technologies [2]. Assistive computer vision (ACV) is one of the growing research areas [3], which can be used to achieve navigation and object recognition. Among them, navigation can be defined as a guided collision-free movement toward the desired target, while object recognition can be defined as the semantic analysis of specific objects and the surrounding environment. The challenge is how to achieve real-time and accurate navigation and object recognition in unstructured environments. Previous works on assistive systems separately investigated navigation [4]–[7] or target recognition [8]–[11]. However, few studies jointly consider these functions to provide the BVI with comprehensive environmental perception. Based on the above reasons, it is necessary and urgent to develop comprehensive wearable assistance systems for the BVI.

The complexity of the indoor environment makes the application of ACV in the indoor environment a challenging task. This article develops a wearable assistance cognitive system that can help indoor sensing and navigation for the visually

Manuscript received 26 July 2021; revised 8 December 2021; accepted 21 January 2022. Date of publication 28 January 2022; date of current version 13 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61625303, Grant U1913601, and Grant U2013601; in part by the Anhui Provincial Natural Science Foundation; in part by the Anhui Energy-Internet Joint Program Grant 2008085UD01; and in part by the National Key Research and Development Program of China under Grant 2018YFC2001602. *(Corresponding author: Zhijun Li.)*

Approval of all ethical and experimental procedures and protocols was granted by Ethics Committee of Yueyang Hospital of integrated traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine under Application No. ChiCTR2000031162.

Guoxin Li, Jiaqi Xu, and Zhijun Li are with the Department of Automation, University of Science and Technology of China, Hefei 230026, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230031, China (e-mail: zjli@ieee.org).

Chao Chen is with the Institute of Advanced Technology, University of Science and Technology of China, Hefei 230031, China.

Zhen Kan is with the Department of Automation, University of Science and Technology of China, Hefei 230026, China (e-mail: zkan@ustc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2022.3146828.

Digital Object Identifier 10.1109/TCDS.2022.3146828

impaired. This system can provide effective information for the BVI, and improve the perception ability and mobility. The overview of the system is shown in Fig. 1, where the images obtained by an RGB-D camera are encoded and transmitted to the remote server through the network. Depending on the mode selected by the user, the server processes the images and sends the results to the embedded board, which then is transmitted to the BVI via verbal feedback. The haptic modules on the BVI's legs can detect the front obstacles and provide vibration feedback determined by the distance to the obstacle. The contributions are summarized as follows.

1) A wearable system, which is built upon low cost devices, is developed to improve the perception of the environment for the BVI.
2) An indoor navigation algorithm based on real-time positioning and wayfinding is developed, which can realize positioning and indoor navigation through simultaneous localization and mapping (SLAM) and fuzzy-based navigation function.
3) A multitarget recognition algorithm based on deep neural networks is proposed, which can realize real-time and accurate object detection, face recognition, and text recognition.

II. Related Works

Recent research are devoted to the development of sensor-based technologies to help the visually impaired navigate. Many of these works focus on the development of new technologies to replace visual systems with other sources, such as haptic or auditory information.

In [12], an enhanced obstacle avoidance method was proposed to help the visually impaired avoid obstacles. The problem of obstacle collision was also addressed using an RGB-D camera in [13]. Most of these systems require users to wear heavy and conspicuous cameras, which makes it hard for daily life. Nevertheless, it is proved that visually impaired people can benefit from depth estimation, which helps them walk safely and freely. These works relied on classic vision algorithms to perform SLAM and to detect obstacles in the surroundings [14]–[18]. Three-dimensional (3-D) cloud information was obtained through dual camera or RGB-D camera, the inertial measurement unit (IMU) was used to provide initial direction [17], and the keyframe matching was used to find the location of the BVI in an indoor environment [18]. The location and target point coordinates were used for path planning.

One of the most important issues for BVI in daily life is object recognition. Traditional methods do not perform well in an unstructured environment. Recently, deep learning technologies are being increasingly used in visual aids [19]–[25]. For example, a deep learning method for object recognition was proposed [21]. However, due to the complexity of buildings in different scenarios, the realization of this technology is very difficult. In [22], a wearable visual aid device based on deep learning technology and RGB-D depth camera was proposed, which can classify the detected obstacles. In [23], a smartphone-based navigation system was proposed, which can detect obstacles in front of the user through an object detection algorithm. The distance to the obstacles is calculated based on pixel density, focal length, and camera height. However, such method is not accurate, and is sensitive to the error in distance estimation. In [25], a wearable blind-assistive device can help BVI read by using microcameras mounted on their fingers and help them find the correct reading layout through tactile and auditory feedback.

When objects are recognized, image captioning is essential for users to help them understand the surrounding objects. Image captioning is an image-to-sequence that takes pixels as input and outputs a sequence of words or subwords by visual encoding and the language model. In [26], image description as the seminal approach by activating last layers of the CNN was proposed, where the output of a GoogleNet pretrained on ImageNet was fed to the initial hidden state of the language model. Similarly, in [27], global features extracted from AlexNet were used as the input for a language model. Global CNN features have been employed for a large variety of image captioning models by their simplicity and compactness of representation [28]–[30]. However, this paradigm also leads to excessive compression of information and lacks granularity. The approaches in [31] and [32] introduced additive attention mechanisms that have increased the granularity level of visual encoding. In [33], a novel abstract scene graph model was developed to describe the object in a fine-grained way, focusing on the content of interest to the user. These approaches almost adopt the same model (i.e., Faster R-CNN trained on visual genome) as the object detector backbone, showing its remarkable performance.

Despite the above works, there are challenges and limitations in existing methods. Most methods only implement a single function, which makes it difficult to be used in the daily life of BVI. As a wearable device is designed, balancing the reliability and real-time performance of the system and excellent human–computer interaction are also worthy of attention. Most systems have demonstrated that the use of audio and haptic feedback helps to transmit information to users [25], [34].

III. Overall Framework

The wearable assistance system takes the feedback from the wearable visual device using the RGB-D camera as input and describes the objects in front of the user in detail according to the user's needs, so the user can act correspondingly and avoid collision or dangerous situations. The environment information is collected and continuously transmitted to the user through verbal feedback and haptic feedback.

A. System Overview

The framework of the system shown in Fig. 2 consists of the wearable visual device, the embedded computer, and the remote serve. The wearable visual device can read the color image and the depth image in real time by an RGB-D camera and transfers these two images to the embedded computer. In the embedded computer, we align the images with

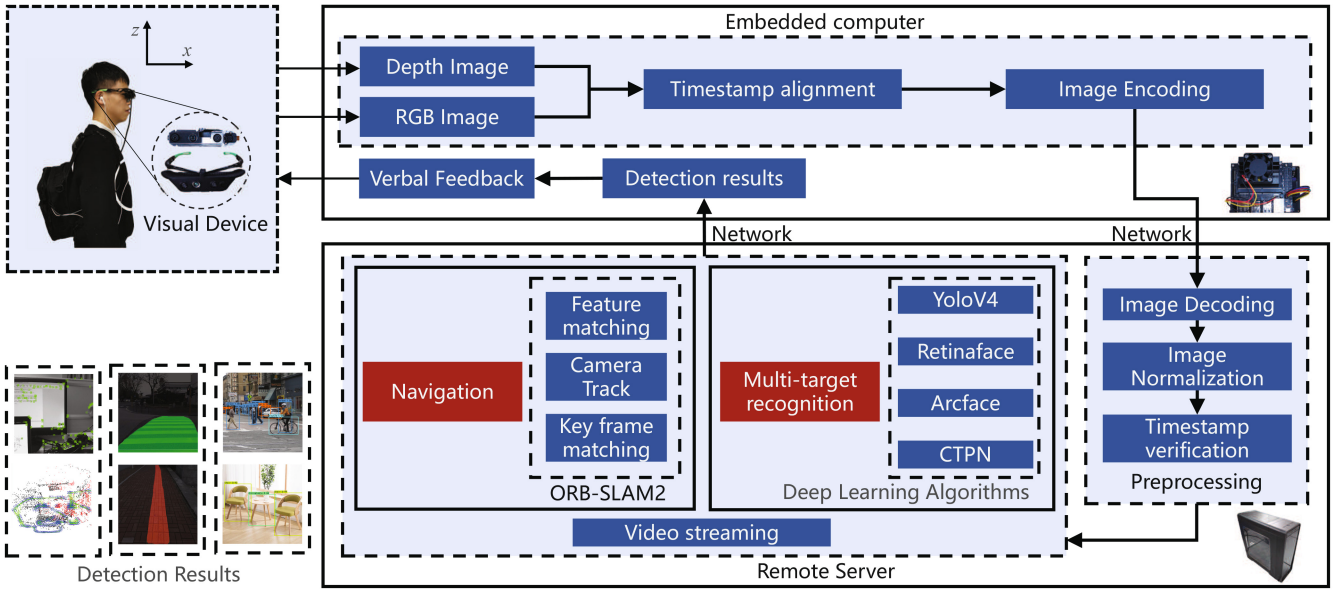


Fig. 2. Framework of the wearable assistance system.

the timestamp and encode the images. Then, the images are transmitted to the remote server through the network, where those images are decoded. We compose the decoded images into a video stream. Based on the video stream, two parallel algorithms, including the oriented fast and rotated brief (ORB)-SLAM2 algorithm for navigation and artificial intelligence algorithms for multitarget recognition, are applied in four work modes, which consist of indoor navigation, object detection, face recognition, and scene text recognition. The system switches between several work modes through voice wake-up and voice recognition and calls different algorithms to analyze the visual image in different modes. The recognition results would be transmitted to the embedded development board through the network, combined into one sentence, and processed as the audio to the BVI.

An indoor navigation method based on real-time positioning and wayfinding is used to empower navigation capability for the BVI. We use real-time positioning and mapping technology to construct indoor maps for the BVI in advance. When the system is running in navigation mode, SLAM is used for indoor positioning of the BVI. After the indoor positioning is completed, the system uses a wayfinding approach to calculate a safe walking route for the BVI. According to the deviation between the current route of the BVI and the desired one, the system gives users navigation instructions to complete indoor navigation.

In order that the BVI can better perceive the object information in the surrounding environment, the multitarget recognition is a synthesis of object detection, face recognition, and scene text recognitions. These functionalities can help the BVI complete different target recognition, avoid obstacles, or find specific objects. In addition, in order to detect obstacles precisely, haptic modules are attached to the legs of the BVI. The haptic module can detect the obstacles in front of the BVI, and produce the haptic feedback to the BVI through the corresponding vibration.

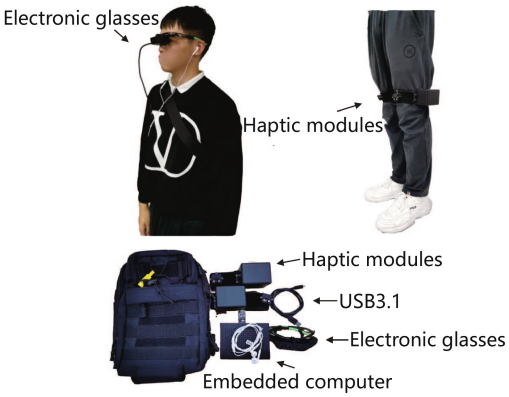


Fig. 3. Hardware of the wearable assistance system.

B. Hardware Setup

As Fig. 3 shows, the wearable assistance system is composed of an embedded computer, electronic glasses, and haptic modules, which costs in total around U.S. $400. The embedded computer is composed of the NVIDIA Jeston Nano 4-GB embedded computer and a 3-D printing shell. The embedded computer is equipped with a battery, which can run for more than 12 h. The price of the embedded board is about U.S. $140.

The electronic glasses are composed of a RealSense D435i depth camera and a 3-D printing shell, as shown in Fig. 4. The depth camera is used to collect RGB images and depth images at a rate of 30 frames/s. The size of the electronic glasses is similar to the ordinary glasses. The 3-D printed shell can prevent the depth camera from being damaged. The electronic glasses' weight is 79 g, which is suitable for people to wear. The price of the glasses is about U.S. $200.

The haptic module is composed of an ultrasonic module (DFRobot and URM09-Analog), a vibration module (YwRobot vibration module), an embedded board, and a 3-D printed shell. The haptic modules are attached with the leg to

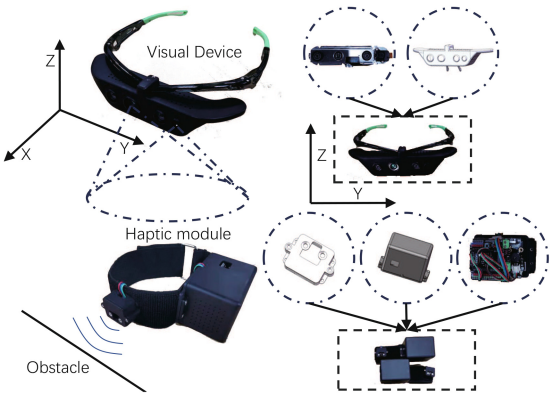


Fig. 4. Electronic glasses and haptic modules.

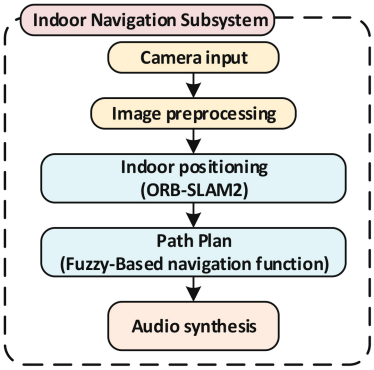


Fig. 5. Structure of indoor navigation subsystem.

assist the user in detecting the front obstacles. The price of haptic modules is about U.S. $60.

The calculation of the system is performed on a remote server. The CPU of the remote server is Intel i7-8700 with 64-GB RAM. In addition, the server is also equipped with an Nvidia GTX1080 for graphics operations. The operating system of the remote server is Ubuntu 18.04.

IV. System Description

This system is constructed from several components, which can be seen from Fig. 2. The functionality of each subsystem is described as follows.

1) The indoor navigation subsystem combines the ORB-SLAM2 algorithm [35] and the enhanced navigation method to perform indoor positioning and navigation. In an unstructured environment, the traditional navigation function is easy to fall into the local optimal solution and the wayfinding is slow. In order to solve this problem, we use a fuzzy logic systems to optimize the navigation function and enhance its searching speed.
2) The multitarget recognition subsystem can provide users with a variety of target recognition capabilities, such as specific obstacle detection, object recognition, facial recognition, and reading capabilities. In order to meet the recommended design criteria for a visual assistance device, we have lightened the deep neural network used in the multitarget detection. To enhance the detection accuracy declined by the lightweight of the network, we have made a lot of structural improvements for object detection.
3) The human–computer interaction subsystem has two functions. One is to switch different working modes according to the user's instruction, and the other is to remind users of the current environment through verbal and haptic feedback based on the recognition results.

In this article, we implement the above-mentioned subsystems with few sensors and feedback devices, thereby reducing the load to the BVI. In order to reduce the cost and power consumption, we utilize the remote server to reduce the local heavy computation and send the instructions back to the local unit.

A. Indoor Navigation

In this subsystem, we propose an indoor navigation based on real-time positioning and wayfinding to provide navigation for the visually impaired. The algorithm combines ORB-SLAM2 and fuzzy-based navigation functions to achieve indoor navigation. The structure of the indoor navigation subsystem is shown in Fig. 5.

ORB-SLAM2 is capable of map reuse, loop closing, and relocalization, where ORB operators are used for feature matching [35]. Compared with traditional operators, such as speed-up robust feature (SURF) [36] and scale-invariant feature transform (SIFT) [37], the operation speed of ORB operators is faster, which is conducive to building real-time systems. The navigation function completes the route search by establishing a potential field map of the current environment [38]. In an unstructured environment, the existence of a large number of obstacles can easily lead to insignificant potential field map gradients, which can easily cause the wayfinding to fall into a local optimal solution and slow down the search speed. In order to solve the shortcomings of the navigation function, this article proposes a navigation function based on the fuzzy logic systems, which adjusts the gradient of the navigation function according to the relative distance of coordinate points, obstacles, and target points. This algorithm can reduce the possibility of falling into a local optimal solution and speed up the path search.

The subsystem can be divided into two steps: 1) indoor positioning and 2) fuzzy-based navigation function used for wayfinding. The parsing of the navigation is shown in Algorithm 1.

1) Indoor Positioning: ORB-SLAM2 is used to obtain 3-D sparse point cloud maps of environments for the indoor positioning. During the SLAM, the user can send real-time instructions to record the location. At the end of the SLAM, the map, camera track, and keyframe information can be automatically saved, and the map can be loaded in the next run to enter the positioning mode. The record, save-map, and load-map methods are integrated into the system.

After the user's positioning is completed based on the 3-D sparse point cloud map, we can plan the path in real time through the fuzzy-based navigation function.

2) Fuzzy-Based Navigation Function: In this section, we construct the obstacle-dependent potential function based on

Algorithm 1: Indoor Navigation Subsystem

input : The video stream, target location
output: The way forward

```
subsys = GetNavigationSubsystem();
while < EXCLAMATION > VideoStream.end() do
    CurrentFrame = VideoStream.NextFrame();
    if FindObstacle(CurrentFrame) then
        subsys.UpdateMap(CurrentFrame);
    end
    if TargetLoc then
        f_{n_d}(p) ← Σ_{1≤i≤n_d} 2^{i-1} τ(p; x_i, y_i);
        IsMatched ← subsys.MatchDes(f_{n_d}(p)) ;
        if IsMatched then
            CurrentLoc ← subsys.GetLocation() ;
            subsys.PathPlan(CurrentLoc, TargetLoc);
        end
    end
    return subsys.WayForward();
end
```

the known position. $r \triangleq (x, y)^T$ is the position of the user. We assume that $r \in \mathbb{X} \triangleq \mathrm{cl}(B_R) \subset \mathbb{R}^2$, where B_R is an open ball of radius R containing the origin and $\mathrm{cl}(\cdot)$ denotes closure. There are some isolated obstacles that can be expressed by the formula $\mathbb{O}_i \in \mathbb{X}$ for $i \in \{1, \ldots, m\}$ in the user's workspace $\mathbb{X}$. For every $i \neq j$ with $i, j \in \{1, \ldots, m\}$, $\mathbb{O}_i \cap \mathbb{O}_j = \emptyset$. The workspace boundary is $\mathbb{O}_0$, which is defined as an obstacle, expressed as the complement of $\mathbb{X}$ in $\mathbb{R}$: $\mathbb{O}_0 \triangleq \mathbb{R}^2 \backslash \mathbb{X}$, and that each obstacle $\mathbb{O}_i (i = 1, \ldots, m)$ can be expressed by a real-valued function β_i as follows:

$$\mathbb{O}_i = \{r \in \mathbb{X} | \beta_i(r) \leq 0\} \tag{1}$$

where $r = (x, y)^T$. The obstacle function β_i is zero on the boundary of $\mathbb{O}_i$, positive outside of $\mathbb{O}_i$, and negative inside of $\mathbb{O}_i$. Then, the multiple obstacles is as $\beta(r) = \prod_{i=1}^{m} \beta_i(r)$.

Suppose that a map reflects the free structure space of the user. It has a minimum value at the target point r_d. For the 2-D space with circular obstacles, we construct the obstacle-dependent potential function based on the known position as follows:

$$\beta_j = \left(x - x_{j0}\right)^2 + \left(y - y_{j0}\right)^2 - \rho_j^2 \quad \mathrm{j} \in \{1, \ldots, m\} \tag{2}$$

where $q_j = (x_{j0}, y_{j0})$ is the coordinate position of the center of the obstacle, ρ_j is the radius, and m is the number of obstacles.

For a 2-D space with circular obstacles, we construct the obstacle-dependent potential function based on the known position as follows [38]:

$$\Phi(r) = \frac{\|r - r_d\|^2}{\left(\|r - r_d\|^{2\kappa} + \beta(r)\right)^{1/\kappa}} \tag{3}$$

where $\kappa > 0$ is a sufficiently large value.

Because there is only a minimum value in formula $\Phi(r)$ at the target point r_d and the maximum value at the boundary of the obstacle, the gradient of $\Phi(r)$ can approach the minimum value without passing through the boundary of the obstacle.

TABLE I
FUZZY RULES FOR FUZZY LOGIC SYSTEM

l_o	l_t			
	ZO	PS	PM	PB
ZO	BE	BE	BE	BE
PS	ME	ME	ME	BE
PM	SE	SE	ME	BE
PB	NE	SE	ME	BE

In order to improve the efficiency of path planning, we hope to increase the potential energy when the user is far away from the target point and reduce the potential energy when the user approaches the target point. To achieve greater gradient descent when planning the path, the parameter α is introduced to the potential energy function, where the value of α is tuned by fuzzy rules.

We consider a dual-input and single-output fuzzy logic system. The dual inputs of the system are the normalized Euclidean distance of the user from the target l_t and the nearest obstacle l_o, and the single output is the adjustment parameter α of the potential energy. The linguistic variables of input and output choose continuous theoretical domain and simple membership function. Based on the normalized Euclidean distance, the theory fields of the linguistic variable l_t and l_o are both set to (0, 1). The theory field of linguistic variable α is set to (1, 2). The triangular membership function in the fuzzy system is chosen as

$$\mu(z, a, b, c) = \begin{cases} 0, & z < a \\ \frac{z-a}{b-a}, & a \leq z \leq b \\ \frac{c-l}{z-b}, & b \leq z \leq c \\ 0, & z > c \end{cases} \tag{4}$$

where $z = l_t, l_o, \alpha$, z is the linguistic variable associated with the inputs and output of the fuzzy system, the parameters a and c represent the feet of the triangle, and b represents the peak of the triangle.

In the theory field (0, 1) of l_t and l_o, the language values of the variables are defined as {ZO, PS, PM, PB}, which denote zero, positive small, positive middle, and positive big, respectively. For linguistic variable α, the language values in the theory field (1, 2) are defined as {none effect (NE), small effect (SE), middle effect (ME), big effect (BE)}. Table I illustrates the nonlinear mapping of the fuzzy rules. The output can be calculated by defuzzification using the center of gravity (COG) as

$$\alpha = \frac{\sum_{i=1}^{j} \mu_{A_i}(\alpha_i)\alpha_i}{\sum_{i=1}^{j} \mu_{A_i}(\alpha_i)} \tag{5}$$

where A_i denotes the fuzzy sets, j is the number of A_i, $\mu_{A_i}(\alpha_i)$ denotes the membership degree of α_i to fuzzy set A_i, and α_i is the base center value for corresponding membership function. Fig. 6 shows the variation curved surface of the input and output of the fuzzy logic system. Then, substituting (3), the fuzzy-based potential function can be rewritten as

$$\varphi(r) = \frac{\alpha\|r - r_d\|^2}{\left(\|r - r_d\|^{2\kappa} + \beta(r)\right)^{1/\kappa}}. \tag{6}$$

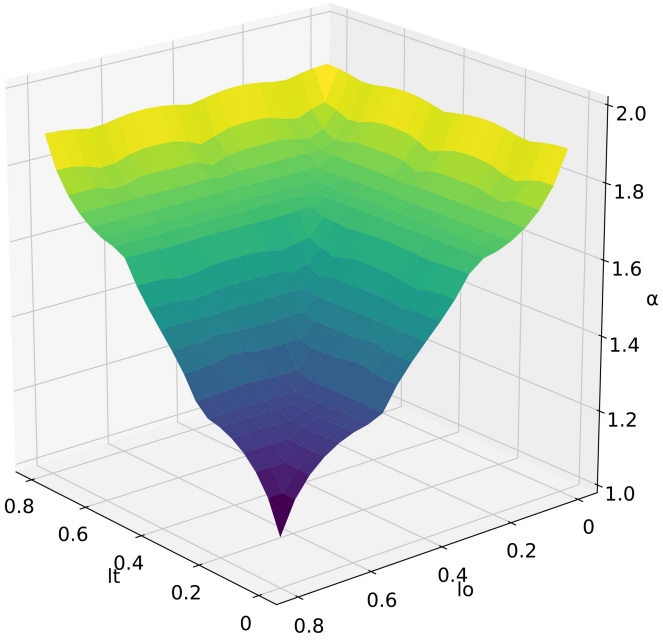


Fig. 6. Variation curved surface of the input and output of fuzzy logic system.

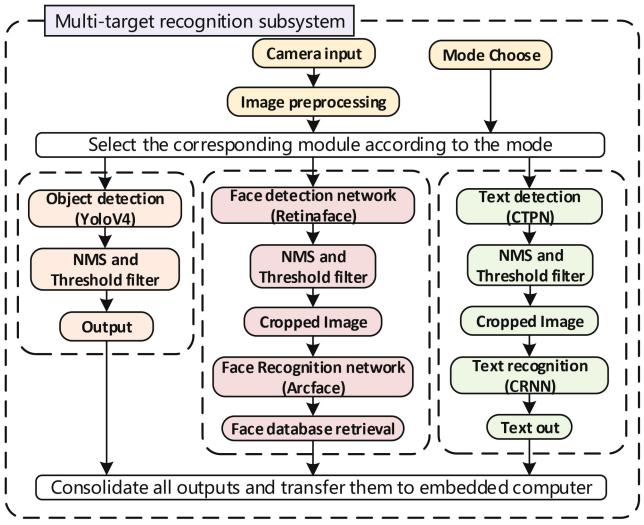


Fig. 7. Structure of multitarget recognition subsystem.

B. Multitarget Recognition

In an unstructured environment, the multitarget detection must meet the recommended design criteria for a visual assistance device, which can prevent BVI collisions and help the system quickly locate the target. In order to meet the system's real-time requirements for the algorithm, we have lightened various deep neural networks in the multitarget detection. Besides, we have made a lot of structural improvements to tackle the dilemma where the accuracy decreased due to the lightweight of the network.

The structure of the multitarget recognition subsystem is shown in Fig. 7, which provides users with multiple target recognition capabilities, including specific obstacle detection, object recognition, face recognition, and reading capabilities.

Algorithm 2: Multitarget Recognition Subsystem

input : The video stream, mode, face mode
output: The detection result

```
subsys = GetRecognitionSubsystem();
yolo = Yolo();
// Initial face recognition model
arcface = Arcface();
face = Retinaface(arcface);
// Initial text recognition model
crnn = CRNN();
ocr = CTPN(crnn);
while !VideoStream.end() do
    CurrentFrame = VideoStream.NextFrame();
    if mode == 1 then
        objResult ← yolo.detectimg(CurrentFrame);
    else if mode == 2 then
        ocrResult ← ocr.detectimg(CurrentFrame);
    end
    if facemode == 1 then
        faceResult ← face.detectimg(CurrentFrame);
    end
    subsys.SortResult(objResult, ocrResult, faceResult);
    return subsys.GetResult();
end
```

The algorithm of the multitarget recognition is outlined in Algorithm 2.

1) Object Detection: The object detection needs to be customized according to the users' demands and takes into account the balance of good processing speed and enough accuracy, which prevent the BVI from accidental collision or injury. Recently, You Only Look Once V4 (YOLOv4) was proposed in 2020 [39], [40]. Although YOLOv4 achieves great detection accuracy and detection speed, the amount of parameters of the network is still very large for wearable devices applied to BVI. We need to optimize the calculation speed based on the YOLOv4 network and maintain excellent detection accuracy.

MobileNetV3 was proposed by Google in 2019, which is a lightweight convolutional neural network used in mobile or embedded devices [41]. MobileNetV3 is a combination of the features of previous different versions of mobilenet and introduces new technologies. The depthwise separable convolution is used to reduce the amount of parameters in the model. The inverted residual with linear bottleneck is used to reduce the amount of calculation while improving the feature extraction ability of the network. In addition, MobileNetV3 introduced a channel-based attention module to enhance feature extraction capability of the model, and introduces the hard-swish activation function to increase the nonlinear feature extraction capability of the network.

We divide the neural network structure of YOLOv4 into Backbone, SPP, Panet, and Yolo Head. The parameters of YOLOv4 are mainly concentrated in Backbone and Panet, so we modify these two parts. The function of Backbone is

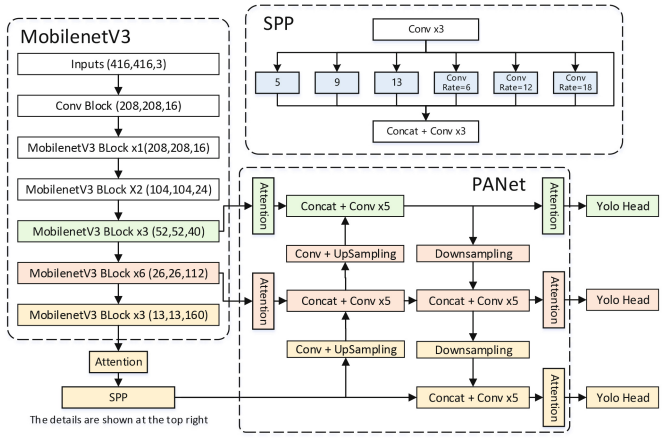


Fig. 8. Network structure of MobileNetV3-YOLOv4-Lite.

to provide preliminary feature extraction for neural networks. The original Backbone of YOLOv4 is CSP-darknet53. We will replace it with MobileNetV3. This replacement can reduce 20 million parameters for YOLOv4, but it still has 40 million parameters. Therefore, we still need to modify Panet. We replace the traditional 3×3 in Panet with the depthwise separable convolutions. This replacement can reduce 30 million parameters for YOLOv4. With the above modification, YOLOv4 only has 10 million parameters left. In order to improve the detection performance of the network, we add attention mechanisms to the Panet of YOLOv4. In order to improve the receptive field of the neural network, we added dilated convolution to the SPP of YOLOv4. We named the modified object detection algorithm as MobileNetV3-YOLOv4-Lite. Compared with the original YOLOv4, MobileNetV3-YOLOv4-Lite significantly reduces the number of parameters. The modified network structure is shown in Fig. 8.

2) Face Recognition: Many high-precision face detection and face recognition often have the shortcomings of slow detection speed and large model, which is not applicable in real-time operation. Locating the face properly is essential to improve the accuracy of face recognition. As a one-stage face detection algorithm, the Retinaface network has a high detection speed while ensuring accuracy [42]. This network implements state of the art on the Wider-Face data sets [43]. To achieve the lightweight Retinaface network, we changed its backbone network to MobileNetV1 [44], which greatly reduces the reasoning time of the face detection network.

In recent years, the face recognition network has been greatly developed. Arcface is a recently proposed network for face recognition [45]. The original Arcface uses the improved LResnet50 as the backbone network [45], leading to a long reasoning time of the network due to its large network parameters. Thus, we also modify its backbone to MobileNetV1 in the face recognition network.

3) Scene Text Recognition: Scene text recognition is generally divided into detection and recognition. The goal of the former is to find the area where the text is as accurate as possible from the picture, while the latter is to recognize the single character in the region on the basis of the former. In natural scenes, text detection is a very difficult task due to the influence of complex backgrounds. Using deep learning algorithms to achieve text detection and recognition can greatly improve the accuracy of detection and recognition.

Connectist text proposal network (CTPN) is a text detection published in 2016 [46]. By combining the advantages of a convolutional neural network and long short-term memory, CTPN can effectively detect the horizontal distribution of text in complex scenes. It is a widely used text detection algorithm nowadays. We can get the text position in the picture through CTPN.

After obtaining the position of the text, we need to recognize the text. The convolutional recurrent neural network (CRNN) is proposed to integrate feature extraction, sequence modeling, and transcription into a unified framework [47]. It combines the advantages of a convolutional neural network and recurrent neural network to realize end-to-end character recognition. The combination of these two deep neural networks can greatly improve the accuracy of scene text recognition.

C. Human–Computer Interaction

When perceiving the surrounding scene (including objects, people, and text) or navigating indoors, the system needs to inform users of appropriate information and guidance. The audio interaction and haptic modules are used together to provide required cognitive information in our system, which ensures efficient and real-time human–computer interaction. The audio performs intuitive sound interaction through headphones, while the haptic module transmits proprioceptive feedback to the user through vibration on the user's legs.

1) Audio Interaction: Audio interaction includes the broadcasting of detected information and the user's voice command input. On the one hand, the embedded computer first combines the text according to the detection results of the current images, and summarize all the detection results into a sentence. After completing the sentence combination, the embedded computer uses the text-to-speech engine for audio synthesis. The audio eventually transmits to the user through the headphone, including object orientation and semantic information. On the other hand, this headphone can be used for voice input to realize the control of the system as well.

All functions of the above subsystems can be run and determined by the user's voice instruction, as shown in Fig. 9. When the instruction is received by the system, the mode will start running or switch, and the detection results from the corresponding subsystem will be fed back to the user.

2) Haptic Modules: Haptic modules provide vibration feedback according to the detection results of the ultrasonic module. When there is an obstacle (within the preset range of one meter) in front of the user, the vibration module on the legs starts to vibrate. As the user approaches the obstacle, the vibration intensity becomes stronger. The relationship between vibration intensity and obstacle distance is as follows:

$$\text{intensity} = \frac{\exp(-d_{\text{obstacle}})}{1 - \exp(-d_{\text{obstacle}})} \tag{7}$$

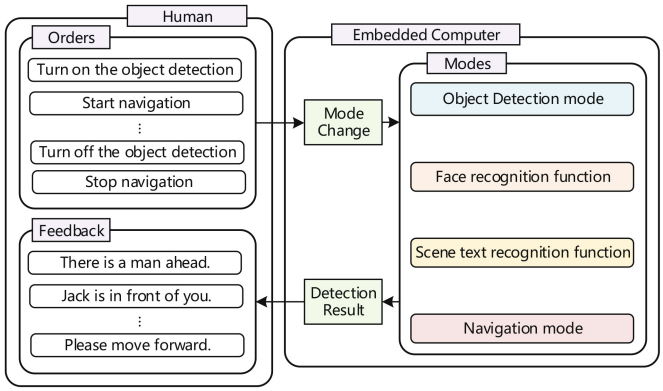


Fig. 9. Audio interaction between human and system.

where d_{obstacle} represents the obstacle distance detected by the haptic module, and intensity represents the vibration intensity of the vibration module, with a value between $0-1$.

V. Experiments and Results

A. Accuracy of Experiment

During the training of the model, we divide the data set used into the training set, validation set, and test set. By simulating on the test set, we can evaluate the current model on the test set.

For the object detection, the metric for mean average precision (mAP) is used for evaluating the model performance. In the object detection model, the confidence threshold is chosen as 0.5. We choose F_1 as a comprehensive metric of recall and precision under the confidence threshold of 0.5, which is defined as

$$F_1 = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{8}$$

We selected 25 000 images as the object detection data set, in which the ratio of the training set, validation set, and test set is 8:1:1. These images are obtained in the daily life or collected from the Web site. Then, the data set is annotated manually. Each image includes one or more objects with different backgrounds. We choose mAP as the performance metric of the object detection algorithm, which conforms to the operation rules of Pascal VOC Challenge 2010 [48]. The network is trained 100 epochs using the Adam optimizer. The initial learning rate lr_{init} is 10^{-3}. The learning rate decreases with the number of training epochs and follows an exponential function. The learning rate lr can be calculated as

$$\text{lr} = \text{lr}_{\text{init}} \times 0.95^{\text{epoch}_n} \tag{9}$$

where epoch_n is the number of iterations of the current epoch.

The object detection accuracy is shown in Table II, and the loss of MobileNetV3-YOLOv4-Lite in the training process is shown in Fig. 10.

Five different object detection algorithms are selected for comparison. Their running speed can be compared using frames per second (FPS), and the detection accuracy can be compared using mAP. The comparison experiment results are shown in Table III.

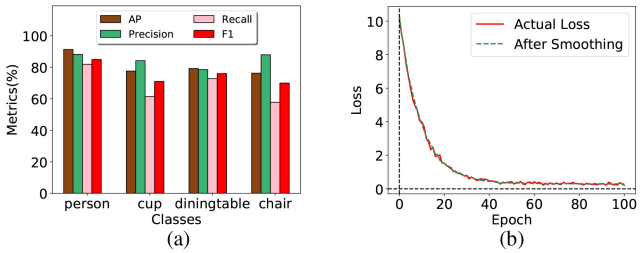


Fig. 10. (a) Metrics of MobileNetV3-YOLOv4-Lite. (b) Loss of MobileNetV3-YOLOv4-Lite in training process.

TABLE II
MAP OF MOBILENETV3-YOLOV4-LITE IN OBJECT DECTION

metrics	person	cup	diningtable	chair	Average
AP (%)	91.31	77.61	79.23	76.27	81.11
Precision (%)	88.19	84.21	78.53	87.93	84.72
Recall (%)	81.93	61.41	72.82	57.80	68.49
F_1	0.85	0.71	0.76	0.70	0.76

TABLE III
MAP AND FPS OF DIFFERENT ALGORITHMS

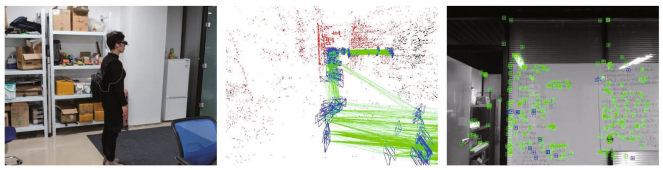
methods	mAP	Precision	Recall	F_1	FPS
SSD [49]	76.89	82.98	58.77	0.69	37
Retinanet [50]	81.95	85.25	72.93	0.79	11
Resnet50-Faster R-CNN [51]	76.86	67.69	81.56	0.74	3
CSPDarkNet53-YOLOv4	85.67	88.93	74.54	0.81	14
MobileNetV3-YOLOv4-Lite	81.11	84.72	68.49	0.76	39

Compared with the original YOLOv4, our proposed MobileNetV3-YOLOv4-Lite only reduces mAP by 5.32%, but improves FPS by 178.57%. Compared with Retinanet, our proposed MobileNetV3-YOLOv4-Lite only reduces mAP by 1.03%, but improves FPS by 254.55%. Compared with SSD and Faster R-CNN, MobileNetV3-YOLOv4-Lite achieves great improvements in detection speed and detection accuracy. Through the comparison of FPS and mAP in different networks, we can conclude that MobileNetV3-YOLOv4-Lite has achieved a great balance between speed and accuracy.

B. User Experiment

In experiments, two BVI volunteers (User1 and User2) and three healthy subjects (User3, User4, and User5) are invited to wear the assistance system to carry out the experiment. These experiments were conducted in the Wearable Robotics and Autonomous Unmanned Systems Laboratory at the University of Science and Technology of China. All procedures of this study were approved by the Ethics Committee of Yueyang Hospital of integrated traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine.

We simulated five daily activities, namely, "Navigation," "Find a chair," "Find a cup," "Face recognition," and "Book reading." The specific requirements for these five daily activities can be found below. Both BVI volunteers participated in all tasks, while three healthy subjects involved three tasks among them.



(a) (b) (c)

Fig. 11. (a) User is using the system for navigation. (b) Corresponding 3-D sparse point cloud map. (c) Matching of ORB feature.

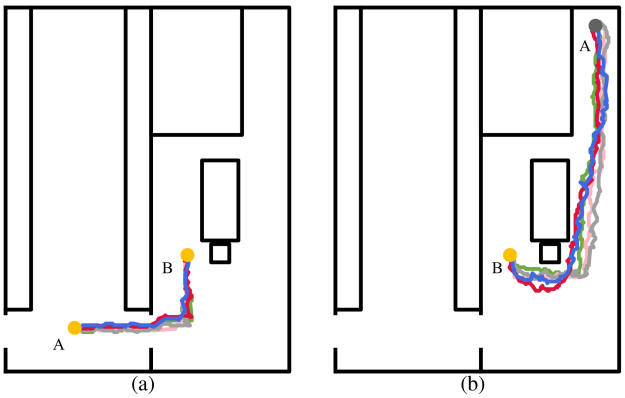


Fig. 12. Walking tracks of "Navigation." The user needs to travel from start location A to end location B according to instructions. The subplot (a) Path I and (b) Path II.

TABLE IV
USER PERFORMANCE IN NAVIGATION

Users	User1	User2	Average
Path I			
No. of Trials	5	3	4
Speed (m/s)	0.13	0.15	0.14
Success rate(%)	100.00 (5/5)	100.00 (3/3)	100.00
Path II			
No. of Trials	5	3	4
Speed (m/s)	0.15	0.12	0.135
Success rate(%)	100.00 (5/5)	66.67 (2/3)	83.33
Note	(n/N) means that n of N trials are successful		

1) Navigation: Navigation requires the user to walk according to navigation instructions and finally, reach the target point. There are two positions A and B in the environment. The user needs to start from the initial point A and move to the target point B. During the experiment, the system records the user's walking track through the SLAM.

Random samples of this experiment are provided in Fig. 11. The user1's walking tracks of Navigation are shown in Fig. 12. The difference of users' performance of Navigation is recorded in Table IV, including the number of trials, the average velocity of walking, and the success rate of completing the walking path.

2) Find a Chair: Find a chair requires the user to find a chair and sit down. The user needs to find the chair according to the instructions of the system. We conducted extensive systematic tests with users. The user performed the Find a chair task in the environment shown in Fig. 13, with one target chair and complex background environment. We varied the position

Fig. 13. Object localization task. The target is a chair. The user needs to walk to the chair and sit down.

of the target at each trial, to avoid performance bias caused by the learning of target positions over time. The time of the task operation is recorded.

3) Find a Cup: Find a cup requires the user to find a cup for drinking. In a structured environment, there is a cup on the table and the user sits in front of the chair. The user needs to find the cup and grab it according to the instructions of the system. We change the position of the cup in each experiment to avoid performance deviations caused by learning the target position over time. The time of grabing the cup is recorded.

4) Face Recognition: Face recognition requires the system to recognize the face in front of the user in time and remind the user. In a specific environment, the user sits in front of the monitor, and we use the computer to randomly select one face picture from the test data each time and display it on the monitor. We record the accuracy of face recognition based on the recognition results.

5) Book Reading: Book reading requires users to complete text reading according to instructions. We randomly provide users with printed text for testing, and the success rate is evaluated based on the recognition results of the printed text. If the recognition accuracy rate of the text on this page is greater than 95%, this experiment is considered to be successful.

In the user experiment, the device was worn on a participant walking indoors. The considered environment is similar to daily life, where the initial and target positions are prespecified. The results of the above tasks in Table VI show that the system can effectively help the visually impaired in indoor navigation, object detection of interest, and face recognition. The average success rates of the Navigation, Find a chair, Find a cup, Face recognition, and Book reading are 91.67%, 83.33%, 91.67%, 99.00%, and 85.00%, respectively. The "Average time" in Tables VI and VII devotes the spending average time of completing the task in trials. The environment information around the user can be accurately transmitted to the user through verbal feedback and haptic feedback.

The verbal feedback of the system changes in real time according to the detected target. For example, the description of the third image sample of the first line in Fig. 14 is: "There is a person in front of you and a chair at the bottom left." In our experiments, the haptic modules are attached to the legs, which can detect the front obstacles with more than 50-cm height (such as indoor chairs and tea tables). Moreover, the 3-D sparse point cloud map based on the ORB-SLAM2 is

TABLE V
COMPARISON BETWEEN THE FUNCTIONS REALIZED BY THE WEARABLE ASSISTANCE SYSTEM PROPOSED BY OTHER RESEARCHERS AND THIS WEARABLE ASSISTANCE SYSTEM

Researches	Navigation	Obstacle dection	Object recognition	Book reading	Face recognition
B. Li et al [4]	✓	✓	×	×	×
I. Abu Doush et al [5]	✓	✓	×	×	×
Y. Tao et al [6]	✓	✓	×	×	×
R. Kessler et al [7]	✓	×	×	×	×
M. A. Khan et al [8]	×	✓	✓	✓	×
R. Jafri et al [9]	×	✓	×	×	×
A. Fernndez et al [10]	×	×	×	×	✓
X. Yang et al [11]	×	×	✓	×	×
Mun-Cheon et al [12]	×	✓	×	×	×
ours	✓	✓	✓	✓	✓

TABLE VI
USER PERFORMANCE IN FIND A CHAIR, FIND A CUP, FACE RECOGNITION, AND BOOK READING

Users	User1	User2	Average
Find a Chair			
No. of Trials	15	15	15
Average time(s)	45	53	49
Success rate(%)	86.67 (13/15)	80.00 (12/15)	83.33
Find a Cup			
No. of Trials	12	12	12
Average time(s)	27	24	25.5
Success rate(%)	100.00 (12/12)	83.33 (10/12)	91.67
Face recognition			
No. of Trials	50	50	50
Success rate(%)	100.00 (50/50)	98.00 (49/50)	99.00
Book reading			
No. of Trials	10	10	10
Success rate(%)	80.00 (8/10)	90.00(9/10)	85.00

TABLE VII
USER PERFORMANCE IN NAVIGATION, FACE RECOGNITION, AND BOOK READING

Users	User3	User4	User5	Average
Path III				
No. of Trials	5	4	4	4
Speed (m/s)	0.21	0.19	0.17	0.19
Success rate(%)	100.00 (5/5)	100.00 (4/4)	100.00 (4/4)	100.00
Face recognition				
No. of Trials	49	42	50	49
Success rate(%)	98.00 (48/49)	100.00 (42/42)	100.00 (57/57)	99.00
Book reading				
No. of Trials	9	10	9	9
Success rate(%)	89.00 (8/9)	90.00 (9/10)	100.00 (9/9)	93.00

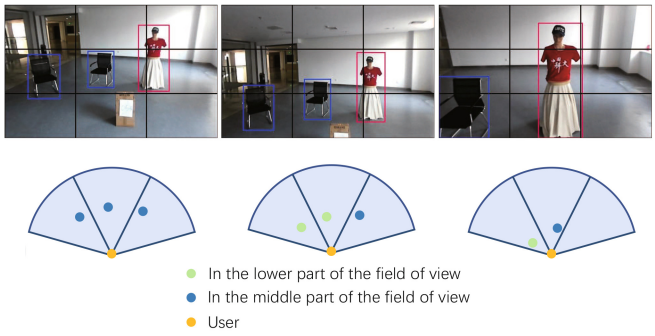


Fig. 14. First line shows the images captured by the camera and the detection results of the object detection algorithm. The second line describes the discrete space corresponding to the image and the relative position of the object. Note that the height difference of the detected obstacles is indicated by different colors.

designed to avoid obstacles. Furthermore, the haptic modules can be easily regulated with the attached places and can detect front obstacles with more than 10-cm height.

Indeed, each task is not completely independent during the experiments. For example, in addition to positioning, mapping, and path planning, navigation also includes target detection and obstacle avoidance. In addition to target detection and recognition, Find a chair also includes navigation. Book reading also includes Face recognition and target detection. Verbal feedback exists in all tasks in real time and provides interaction with the BVI. All tasks driven by subsystems are supervised by the main program in the system, and the current awakened task has the highest priority except obstacle avoidance to ensure the best interaction experience.

In the experiment, the index "Speed" or Average time shows the performance of our system to assist users in performing tasks. The faster the system transmission and processing, the stronger the user's ability to accept instructions and execute activities, and the faster the user's speed or the shorter the completion time. The image processing time of the system is determined by the proposed algorithm. We use the FPS index to measure the real-time performance of the algorithm, as shown in Table III. However, the image transmission time is affected by the network state. We found that the transmission time from the embedded version to the server is about 0.15–0.2 s under the smooth network state, such as 4G or WiFi. While the network is poor at times, the transmission time is up to 1 s. Encouragingly, even, in this case, the users have no obvious discomfort while wearing the device, which shows that the transmission and processing of the system can meet their needs.

To show the comprehensive assistance for multitasks in daily life, we compare the functions of our system with other existing wearable systems, which is shown in Table V.

The pilot experiment shows that our device can fulfill various tasks. Compared with other devices, our devices have more comprehensive functions and can better meet BVI's requirements.

VI. Conclusion

We have proposed a wearable vision-based assistance system, which includes an RGB-D camera, an embedded computer, and two haptic modules. The whole assistance system is composed of three subsystems. The indoor navigation subsystem can fulfill real-time positioning and navigation through the ORB-SLAM2 algorithm. The multitarget recognition subsystem can provide the users with a variety of target recognition capabilities. The human–computer interaction subsystem has two functions. The first is to switch the different working modes according to the user's instruction. The second is to remind the users of the current environment through verbal and vibration feedback. The experiments have verified that the system can work well and satisfy the needs of the BVI in daily use.

Although our developed device has more comprehensive functions compared with the existing devices, it still needs to be further explored. For example, its performance would be influenced by the network status. Future improvements include higher accuracy algorithms with lower latency, and richer and like-human cognitive feedback as well.

References

[1] S. Advani *et al.*, "A multitask grocery assist system for the visually impaired: Smart glasses, gloves, and shopping carts provide auditory and tactile feedback," *IEEE Consum. Electron. Mag.*, vol. 6, no. 1, pp. 73–81, Jan. 2017.

[2] G. M. Farinella, T. Kanade, M. Leo, G. G. Medioni, and M. Trivedi, "Special issue on assistive computer vision and robotics–Part I," *Comput. Vis. Image Understand.*, vol. 148, pp. 1–2, Jul. 2016.

[3] Y.-Z. Hsieh, S.-S. Lin, and F.-X. Xu, "Development of a wearable guide device based on convolutional neural network for blind or visually impaired persons," *Multimedia Tools Appl.*, vol. 79, no. 39, pp. 29473–29491, 2020.

[4] B. Li *et al.*s, "Vision-based mobile indoor assistive navigation aid for blind people," *IEEE Trans. Mobile Comput.*, vol. 18, no. 3, pp. 702–714, Mar. 2019.

[5] I. Abu Doush, S. Alshatnawi, A.-K. Al-Tamimi, B. Alhasan, and S. Hamasha, "ISAB: Integrated indoor navigation system for the blind," *Interact. Comput.*, vol. 29, no. 2, pp. 181–202, Mar. 2017.

[6] Y. Tao and A. Ganz, "Validation and optimization framework for indoor navigation systems using user comments in spatial-temporal context," *IEEE Access*, vol. 7, pp. 159479–159494, 2019.

[7] R. Kessler, M. Bach, and S. P. Heinrich, "Two-tactor vibrotactile navigation information for the blind: Directional resolution and intuitive interpretation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 3, pp. 279–286, Mar. 2017.

[8] M. A. Khan, P. Paul, M. Rashid, M. Hossain, and M. A. R. Ahad, "An AI-based visual aid with integrated reading assistant for the completely blind," *IEEE Trans. HumanMach. Syst.*, vol. 50, no. 6, pp. 507–517, Dec. 2020.

[9] R. Jafri, R. L. Campos, S. A. Ali, and H. R. Arabnia, "Visual and infrared sensor data-based obstacle detection for the visually impaired using the google project tango tablet development kit and the unity engine," *IEEE Access*, vol. 6, pp. 443–454, 2018.

[10] A. Fernández, J. L. Carús, R. Usamentiaga, and R. Casado, "Face recognition and spoofing detection system adapted to visually-impaired people," *IEEE Latin America Trans.*, vol. 14, no. 2, pp. 913–921, Feb. 2016.

[11] X. Yang, S. Yuan, and Y. Tian, "Assistive clothing pattern recognition for visually impaired people," *IEEE Trans. HumanMach. Syst.*, vol. 44, no. 2, pp. 234–243, Apr. 2014.

[12] M.-C. Kang, S.-H. Chae, J.-Y. Sun, S.-H. Lee, and S.-J. Ko, "An enhanced obstacle avoidance method for the visually impaired using deformable grid," *IEEE Trans. Consum. Electron.*, vol. 63, no. 2, pp. 169–177, May 2017.

[13] A. Aladrén, G. López-Nicolás, L. Puig, and J. J. Guerrero, "Navigation assistance for the visually impaired using RGB-D sensor with range expansion," *IEEE Syst. J.*, vol. 10, no. 3, pp. 922–932, Sep. 2016.

[14] S. Li and D. Lee, "RGB-D SLAM in dynamic environments using static point weighting," *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2263–2270, Oct. 2017.

[15] J. A. Hesch and S. I. Roumeliotis, "Design and analysis of a portable indoor localization aid for the visually impaired," *Int. J. Robot. Res.*, vol. 29, no. 11, pp. 1400–1415, 2010.

[16] H. Zhang and C. Ye, "An indoor wayfinding system based on geometric features aided graph SLAM for the visually impaired," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1592–1604, Sep. 2017.

[17] Y. H. Lee and G. Medioni, "Wearable RGB-D indoor navigation system for the blind," in *Proc. Comput. Vis. ECCV Workshops*, 2014, pp. 493–508.

[18] X. Zhang *et al.*, "A SLAM based semantic indoor navigation system for visually impaired users," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2015, pp. 1458–1463.

[19] M. Cornacchia, B. Kakillioglu, Y. Zheng, and S. Velipasalar, "Deep learning-based obstacle detection and classification with portable uncalibrated patterned light," *IEEE Sensors J.*, vol. 18, no. 20, pp. 8416–8425, Oct. 2018.

[20] B. Jiang, J. Yang, Z. Lv, and H. Song, "Wearable vision assistance system based on binocular sensors for visually impaired users," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1375–1383, Apr. 2019.

[21] V. V. Meshram, K. Patil, V. A. Meshram, and F. C. Shu, "An astute assistive device for mobility and object recognition for visually impaired people," *IEEE Trans. HumanMach. Syst.*, vol. 49, no. 5, pp. 449–460, Oct. 2019.

[22] W.-J. Chang, L.-B. Chen, M.-C. Chen, J.-P. Su, C.-Y. Sie, and C.-H. Yang, "Design and implementation of an intelligent assistive system for visually impaired people for aerial obstacle avoidance and fall detection," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10199–10210, Sep. 2020.

[23] B.-S. Lin, C.-C. Lee, and P.-Y. Chiang, "Simple smartphone-based guiding system for visually impaired people," *Sensors*, vol. 17, no. 6, p. 1371, 2017.

[24] S. Panchanathan, S. Chakraborty, and T. McDaniel, "Social interaction assistant: A person-centered approach to enrich social interactions for individuals with visual impairments," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 5, pp. 942–951, Aug. 2016.

[25] Y.-S. Su, C.-H. Chou, Y.-L. Chu, and Z.-Y. Yang, "A finger-worn device for exploring chinese printed text with using CNN algorithm on a micro iot processor," *IEEE Access*, vol. 7, no. 1, pp. 116529–116541, 2019.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164.

[27] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017.

[28] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language CNN for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1231–1240.

[29] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, "GroupCap: Group-based image captioning with structured relevance and diversity constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1345–1353.

[30] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1179–1195.

[31] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2506–2515.

[32] H. Ge, Z. Yan, K. Zhang, M. Zhao, and L. Sun, "Exploring overall contextual information for image captioning in human-like cognitive style," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1754–1763.

[33] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9959–9968.

[34] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 6533–6540.
[35] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.
[36] S. Ehsan, N. Kanwal, A. F. Clark, and K. D. McDonald-Maier, "An algorithm for the contextual adaption of SURF octave selection with good matching performance: Best octaves," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 297–304, Jan. 2012.
[37] T. Chen and L. Chen, "A union matching method for SAR images based on SIFT and edge strength," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4897–4906, Dec. 2014.
[38] E. Rimon and D. E. Koditschek, "Exact robot navigation using artificial potential functions," *IEEE Trans. Robot. Autom.*, vol. 8, no. 5, pp. 501–518, Oct. 1992.
[39] S. Albahli, N. Nida, A. Irtaza, M. H. Yousaf, and M. T. Mahmood, "Melanoma lesion detection and segmentation using YOLOv4-DarkNet and active contour," *IEEE Access*, vol. 8, pp. 198403–198414.
[40] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[41] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1314–1324.
[42] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5202–5211.
[43] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 5525–5533.
[44] Y. Zhou, Y. Liu, G. Han, and Y. Fu, "Face recognition based on the improved MobileNet," in *Proc. IEEE Symp. Series Comput. Intell. (SSCI)*, 2019, pp. 2776–2781.
[45] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4690–4699.
[46] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 56–72.
[47] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.
[48] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[49] J. Li, Q. Hou, J. Xing, and J. Ju, "SSD object detection model based on multi-frequency feature theory," *IEEE Access*, vol. 8, pp. 82294–82305, 2020.
[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

Guoxin Li received the B.S. degree in mechanical engineering and automation from Hefei University of Technology, Hefei, China, in 2016. He is currently pursuing the Ph.D. degree in control science and engineering with the University of Science and Technology of China, Hefei.

His current research interests include human–robot interaction, multitarget recognition, and wearable robotics systems.

Jiaqi Xu received the B.S. degree in automation from Hunan University, Changsha, China, in 2019. He is currently pursuing the M.S. degree in control engineering with the University of Science and Technology of China, Hefei, China.

His current research interests include multitarget recognition, human–computer interaction, and deep learning.

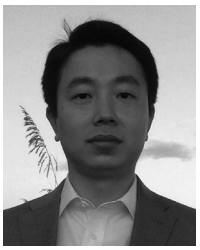
Zhijun Li (Fellow, IEEE) received the Ph.D. degree in mechatronics from Shanghai Jiao Tong University, Shanghai, China, in 2002.

From 2003 to 2005, he was a Postdoctoral Fellow with the Department of Mechanical Engineering and Intelligent Systems, the University of Electro-Communications, Tokyo, Japan. From 2005 to 2006, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and Nanyang Technological University, Singapore. Since 2017, he has been a Professor with the Department of Automation, University of Science and Technology of China, Hefei, China, where he was the Vice Dean of the School of Information Science and Technology in 2019. His current research interests include wearable robotics, teleoperation systems, nonlinear control, and neural-network optimization.

Prof. Li has been the Co-Chair of IEEE SMC Technical Committee on Bio-Mechatronics and Bio-Robotics Systems and IEEE RAS Technical Committee on Neuro-Robotics Systems. He is serving as an Editor-at-Large for *Journal of Intelligent and Robotic Systems* and an associate editor for several IEEE Transactions.

Chao Chen received the B.S. degree in automation from Anhui University, Hefei, China, in 2020. He is currently pursuing the M.S. degree in control engineering with the University of Science and Technology of China, Hefei.

His current research interests include multimodal fusion, human–computer interaction, and deep learning.



Zhen Kan (Member, IEEE) received the Ph.D. degree in mechanical and aerospace engineering from the University of Florida, Gainesville, FL, USA, in 2011.

He was a Postdoctoral Research Fellow with Air Force Research Laboratory, Eglin AFB, FL, USA, and the University of Florida Research and Engineering Education Facility, Shalimar, FL, USA, from 2012 to 2016, and was an Assistant Professor with the Department of Mechanical Engineering, University of Iowa, Iowa City, IA, USA, from 2016 to 2019. He is currently a Professor with the Department of Automation, University of Science and Technology of China, Hefei, China. His research interests include networked control systems, nonlinear control, formal methods, and robotics.

Prof. Kan currently serves on program committees of several internationally recognized scientific and engineering conferences and he is an Associate Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL.