LiSee: A Headphone that Provides All-day Assistance for Blind and Low-vision Users to Reach Surrounding Objects

KAIXIN CHEN, Shenzhen University, China
YONGZHI HUANG, Shenzhen University and Hong Kong University of Science and Technology, China
YICONG CHEN, Shenzhen University, China
HAOBIN ZHONG, Shenzhen University, China
LIHUA LIN, Shenzhen University, China
LU WANG, Shenzhen University, China
KAISHUN WU, Shenzhen University, China

Reaching surrounding target objects is difficult for blind and low-vision (BLV) users, affecting their daily life. Based on interviews and exchanges, we propose an unobtrusive wearable system called LiSee to provide BLV users with all-day assistance. Following a user-centered design method, we carefully designed the LiSee prototype, which integrates various electronic components and is disguised as a neckband headphone such that it is an extension of the existing headphone. The top-level software includes a series of seamless image processing algorithms to solve the challenges brought by the unconstrained wearable form so as to ensure excellent real-time performance. Moreover, users are provided with a personalized guidance scheme so that they can use LiSee quickly based on their personal expertise. Finally, a system evaluation and a user study were completed in the laboratory and participants' homes. The results show that LiSee works robustly, indicating that it can meet the daily needs of most participants to reach surrounding objects.

CCS Concepts: • **Human-centered computing** → **Accessibility systems and tools**; **Sound-based input / output**; • **Computer systems organization** → **Robotics**.

Additional Key Words and Phrases: Reach Object, Wearable System, Speech Interface, Visual Impairments, All-day Assistance

ACM Reference Format:
Kaixin Chen, Yongzhi Huang, Yicong Chen, Haobin Zhong, Lihua Lin, Lu Wang, and Kaishun Wu. 2022. LiSee: A Headphone that Provides All-day Assistance for Blind and Low-vision Users to Reach Surrounding Objects. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 104 (September 2022), 30 pages. https://doi.org/10.1145/3550282

1 INTRODUCTION

Blind and low-vision (BLV) users lack the ability to perceive external objects owing to the lack of vision, and this visual information gap causes difficulties in daily life. In the past few decades, many auxiliary systems have been developed to provide assistance to BLV users. Most of these studies [5, 8, 11–13] have focused on text recognition, object recognition, or navigation. However, BLV users often need to access their surrounding space in daily life,

Authors' addresses: Kaixin Chen, 2017133035@email.szu.edu.cn, Shenzhen University, China; Yongzhi Huang, huangyongzhi@email.szu.edu.cn, Shenzhen University and Hong Kong University of Science and Technology, China; Yicong Chen, 2018151032@email.szu.edu.cn, Shenzhen University, China; Haobin Zhong, 2019281010@email.szu.edu.cn, Shenzhen University, China; Lihua Lin, 2018063031@email.szu.edu.cn, Shenzhen University, China; Lu Wang, wanglu@szu.edu.cn, Shenzhen University, China; Kaishun Wu, wanglu@szu.edu.cn, Shenzhen University, China.

2474-9567/2022/9-ART104 $15.00
https://doi.org/10.1145/3550282

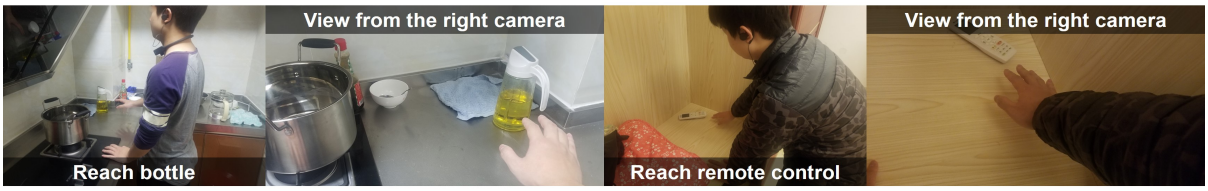


Fig. 1. Our wearable system integrates a binocular camera, microphone and speaker, as well as motherboards, which is an extension of the existing neckband headphone and provides all-day assistance for the BLV user to reach surrounding objects.

that is, they need to reach objects around them. For instance, they may need to reach a water cup from the table, grab a mobile phone, and pick up fallen objects, all of which are important basic living abilities for BLV users. However, as BLV users lack visual cognition and positional awareness, it is difficult for them to complete these tasks. They can only perceive the surrounding environment by hand, but this tactile search is time-consuming and inaccurate. At present, BLV users often rely on the help of those with normal vision, such as family [38], or use mobile apps to seek remote assistance from online workers with normal vision. There are several well-known apps, namely, VizWiz [21] and BeMyEyes [4], that allow BLV users to share photos or real-time videos with workers with normal vision on the platform and ask questions. However, these methods are expensive and sharing real-time videos has caused privacy concerns [15]. In order to overcome these challenges, wearable systems have been proposed that do not require manual assistance for BLV users to reach their objects. The wearable research community has proposed FingerSight [70], Third Eye [85], W-ROMA [55], improved W-ROMA [45] and DLWV2 [72], which are worn on the hand or wrist and provide guidance through the vibration of vibro-motors in different positions. GuideCopter [42] is connected to the fingers of BLV users with a wire, and it uses an unmanned aerial vehicle (UAV) to pull BLV users to reach objects. Two applications based on mobile phones have been developed that can tell BLV users the location of the target object through speech [47, 74]. Unfortunately, these systems require BLV users to wear hand-worn devices (such as gloves and wristbands) or use handheld mobile phones, which prevent BLV users from freely using their hands and thus restricting them from completing other tasks. Reaching surrounding objects may be necessary for BLV users at any time in daily life, and it is accompanied by other tasks. Undoubtedly, devices that are inconvenient to wear and take off are not suitable for frequent use in daily life.

Therefore, we propose the first neckband headphone wearable system, a lossless wearable form called LiSee[1] (Figure 1) to help BLV users reach surrounding objects. The present study was conducted to overcome three key challenges. First, there is no preliminary interview research on BLV users reaching surrounding objects, so it is difficult for researchers to understand the difficulties and requirements of BLV users, which leads to challenges when building a system. Second, how to build a wearable form suitable for BLV users to wear all day and that is frequently used in daily life has not been explored. Third, it is challenging to ensure the robustness of the system in the new unconstrained wearable form. In order to tackle these challenges, we used a user-centered design method and analyzed the needs and habits of BLV users through different types of interviews and experiments. We designed a novel neckband headphone shape, taking the function of the system as the expansion and extension of the existing headphone, which is suitable for BLV users to wear all day and use at any time in their daily life. We also designed a set of seamlessly coupled image processing algorithms to meet the task requirements in the new wearable form, and used cloud and fog collaborative computing to meet the computational power requirements. The research results in the laboratory and participants' homes show that our system could meet

[1]LiSee is a combination of "Listen" and "See". It is also a transliteration of the Chinese character "Lingxi", which means understanding and emotional resonance.

the needs of BLV users to reach objects in their daily life. The contributions of this study can be summarized as follows:

- LiSee is a novel technology using a neckband headphone to assist BLV users reach surrounding objects, which is suitable for BLV users to wear for a long time and use at any time. In addition, the well-designed and inconspicuous system alleviates the problems of esthetics and stigma.
- We followed a user-centered design method and examined the requirements and habits of BLV users through different forms of interviews and experimental research. We jointly designed a set of guidance schemes with BLV users to guide them to reach objects efficiently and naturally. We believe that quantitative and transparent analysis is conducive to the further research and development of this subject.
- We designed a complete set of image processing algorithms to solve the challenges brought by the unconstrained form. At the same time, we deployed the functional modules of the algorithms in fog devices and clouds, which makes up for the lack of computing power of the headphone and provides users with a reliable and smooth user experience.
- We carried out a real scene experiment in the laboratory and at BLV users' homes. Quantitative evaluation and user research results show that our system was easy to adopt, effective, and favored by BLV users.

2 RELATED WORK AND BACKGROUND

We proposes a ready-to-use wearable system that is suitable for long-time wearing to help BLV users reach objects. First, we introduce the common auxiliary systems to help BLV users obtain visual information. Then we introduce the systems that assist BLV users to reach target objects, and discuss their advantages and disadvantages in detail.

2.1 System for Assisting BLV Users to Obtain Visual Information

In order to help BLV users perceive the surrounding environment, researchers have proposed some auxiliary systems. Based on the artificial assistance method, workers with normal vision can answer BLV users' questions remotely by watching shared photos or real-time videos. BeMyEye [4] and VizWiz [21] are two well-known apps. However, sharing photos or real-time videos with strangers has caused privacy concerns [15]. In addition, these applications are expensive [4]. The method based on computer vision can greatly alleviate these problems. Envision AI [5], Seeing AI [13], and LookTel Recognizer [12] are apps that can realize object recognition, color recognition, currency recognition, scene description, and other functions. OrCam [8] smart glasses can describe objects to BLV users. GIST [48] is a gesture interface worn on the chest, which allows BLV users to use different gestures to access the visual information of the space. Fingerreader2.0 [22] places a camera on the BLV user's index finger and helps them recognize characters in a shopping scene. AiSee [23] is a bone conduction headset integrated with a camera that can identify the category and text information of goods. However, these systems can only provide limited help, because they all assume that BLV users can find the target object or just broadcast the visual information flow blindly, and they cannot help BLV users locate the target objects and guide them to reach them.

2.2 Assistance for BLV Users in Locating and Reaching Target Objects

The above system can work well only when BLV users know where the target object is. Therefore, in order to help BLV users locate and reach target objects, researchers have proposed various methods and systems. FingerSight [70] is a wearable ring with four vibration motors to guide BLV users, but it must be connected to a computer to control the vibration, and only a preliminary feasibility study on manual vibration has been done. Third Eye [85] contains a pair of glasses and gloves (there are four vibro-motors on the glove to guide the BLV users), which can only identify the objects of preregistered templates. W-ROMA [55] is handheld auxiliary robot equipment that uses motor movement to keep the gloved hand aligned with the target object. However, it relies

on traditional visual features, which leads to inaccurate classification. In addition, these systems cannot obtain depth information, which is important information for guiding BLV users to reach the target object. Therefore, improved W-ROMA [45] adds a depth camera to the glove. It uses MobileNet-SSD [40] to identify the target object and uses the inertial measurement unit to estimate the posture of the hand. Finally, it uses an electrical stimulation matrix to transfer the direction information to BLV users. However, using the electrical stimulation matrix is not intuitive, so it takes a long time to learn to feel the pattern of electrical stimulation. DLWV2 [72] is a wrist strap with a monocular camera and five vibration motors, which uses motor vibration at five different positions to provide guidance in different directions. However, systems [45, 55, 70, 72, 85] with sensors worn on the hands have obvious disadvantages: BLV users need to straighten their hands and maintain stability to ensure the camera angle of view, and the images taken by the camera are easily blurred in the process of moving the hand, which leads to an unsatisfactory image processing level in practical use. In addition, wearing gloves [45, 55, 85] also restricts BLV users' tactile perception of objects. GuideCopter [42] is a UAV tether system that is connected to the fingers of BLV users with a wire to pull BLV users to reach objects. However, the UAV method is inconvenient for BLV users [18]. In short, the above wearable systems are not suitable for use outside the laboratory, because the connections with computers and the inconvenient forms render the systems unsuitable for long-time use in daily life. However, reaching surrounding objects is a task that is performed often. In contrast, mobile phones are pervasive. AIGuide [74] is a smartphone-based application that uses Apple's augmented reality framework ARKit [3] to detect objects in 3D space, track them in real time, and guide the hands of visually impaired people with auditory and tactile feedback. MobileNet V2 [69] is used to add recognizable categories to solve the problem of insufficient object categories of AIGuide [47]. However, it is hard to keep the handheld camera stable and this requires both hands. As a result, BLV users have only one hand to do other tasks, which is limiting in many scenarios (e.g., cooking in the kitchen). Existing systems generally suffer from poor wearability or occupy both hands, so they cannot be worn for a long time or be used frequently outside the laboratory. On the contrary, we used a user-centered design method and regarded BLV users as partners to propose a neckband headphone suitable for all-day use.

3 UNDERSTANDING THE DIFFICULTIES AND REQUIREMENTS OF BLV USERS TO REACH TARGET OBJECTS

Existing auxiliary devices are not widely used, usually because designers do not understand the situations faced by BLV users in daily life, which leads to the inability to correctly understand the pain points and needs of BLV users and thus the inability to provide effective auxiliary devices. We adopted a user-centered design method [14] and involved the BLV users in our design process [68]. Although relevant interview studies on BLV users' shopping [23, 81], indoor navigation [43], privacy concerns [15], social interaction [44], and family collaboration [24, 77] have been conducted, there is still no in-depth insight into reaching surrounding target objects. Therefore, in order to better understand the pain points and requirements of BLV users to reach target objects, we first conducted semi-structured interviews [56] with BLV users. The reason for using semi-structured interviews is that we believed that guided and flexible communication would gradually help us understand BLV users more deeply and form thematic views.

3.1 Participants

Five of our researchers and eight BLV users (24–58 years of age, including four males and four females, with an average age of 38.9 [SD = 10.6]; P1–P8 in Table 3) participated in this part of the study. In order to facilitate the analysis, we asked them about their basic personal information. Their detailed demographic information is shown in Table 3. All of the participants were from China. We recruited them from local establishments of the China Disabled Persons' Federation (CDPF), blind massage shops and the Internet. When recruiting participants,

we told them our purpose and desire: to understand them through interviews and to cooperate with them to design a user-friendly and reliable auxiliary system to help them reach target objects in their daily lives. We followed the definition of visual impairment by the World Health Organization [63]: blindness is defined as visual acuity < 0.05, while low vision is defined as 0.05 < visual acuity < 0.3. Among the participants, four were blind (P1, P4, P5, P6) and four were low-vision users. The causes included glaucoma, retinitis pigmentosa, cataract, and optic atrophy. The age of vision loss was between birth and 25 years old. Their careers also varied, including beautician, blind masseur, and barrier-free engineer.

3.2 Procedure

The questions were provided in advance by all of the members of our research team and were based on the researchers' considerations and questions. Interviews with P6 and P8 were conducted in the blind massage shops, interviews with P3 and P5 were conducted online, and other interviews were conducted at the participants' homes. The families of two participants (P4 and P7) attended the interview. We first collected and checked their information and asked them to sign the informed consent form. Then, we conducted interviews according to the following subject questions, and recorded answers at the same time:

(1) *When do you need to reach objects*: Do you often need to reach objects in your daily life (including picking up a cup on the table and picking up objects that accidentally fell)? How often are these tasks necessary? What are usual scenarios? What are usual target objects?

(2) *How do you reach your objects*: Do you need to rely on other people or other technologies? If you reach an object yourself, how do you locate the object? What clues do you use to locate the target? Do you use the rest of your vision? Please describe the process clearly.

(3) *Difficulties encountered*: What is the biggest difficulty you encounter when reaching an object?

(4) *Functional requirements*: What functions do you think the system we jointly design should have to help you overcome the difficulties? Think of speech feedback, vibration, or other feedback methods?

(5) *Design requirements*: What design requirements do you think the auxiliary device should meet in order to be used in daily life? How do you prefer to wear this device?

All interviews lasted 30 to 50 minutes.

3.3 Key Topic Discovery

Through semi-structured interviews and observations, we gained a deeper understanding of the process of BLV users reaching objects. We understood their challenges and their design needs. After transcribing the recordings and organizing the notes, we conducted a thematic analysis [25]. Due to space constraints, we discuss three key topics.

3.3.1 Difficulties in Reaching Target Objects. **The need to reach objects frequently**: Unsurprisingly, all of the participants acknowledged that reaching target objects is a frequent difficulty for them because of the lack of visual cognition and positional awareness. However, reaching objects is important. They encounter this situation several times or even dozens of times a day.

Multiple sources of difficulties: Fallen objects, a vague memory, and small sizes cause difficulties. Everyone mentioned that things falling made the target location unknown. A vague memory and small sizes were mentioned 7 times and 6 times, respectively. For objects whose positions often change, it is not easy to remember the last position (e.g., frequently used water cups, keys, and mobile phones). P7 explained that it is sometimes difficult for his girlfriend to put things back in their original position. P2 said that it is difficult for him to reach object because his table is too messy. P5 mentioned that reaching tiny objects such as keys and tiny headphones is a great challenge.

Reaching target objects on a table or the floor: Target objects are mainly located on a table or on the floor. Everyone mentioned that reaching objects on a table is the most common scenario, in places such as the living room, the restaurant, the study, the kitchen, and the bedroom. Participants mainly reach target objects on the floor because they have fallen (P2, P3, P5, P7, P8).

Many strategies are used to find target objects: Participants use a variety of strategies to find objects, including touch, smell, residual vision, and asking for help from others. All of the participants showed that they mainly use touch to find things. They are used to groping for something on the table and then distinguishing objects according to the shape and material. In addition, smell is helpful to distinguish objects (mentioned 6 times). P2 and P7 used residual vision to find objects. In addition, when they tried to explore but still could not find an object, they would turn to their parents or partners (4 times).

3.3.2 Functional Requirements Proposed by BLV Users. It is necessary to develop a system to help BLV users reach objects. According to the functional requirements of participants, we concluded that the system should have the following functions:

Identify and locate objects accurately and quickly: All of the participants wanted the system to accurately and quickly identify objects. When the system can accurately and quickly identify and locate objects, it can compensate for their lack of visual cognition and positional awareness. Therefore, the system should be able to identify and locate objects robustly.

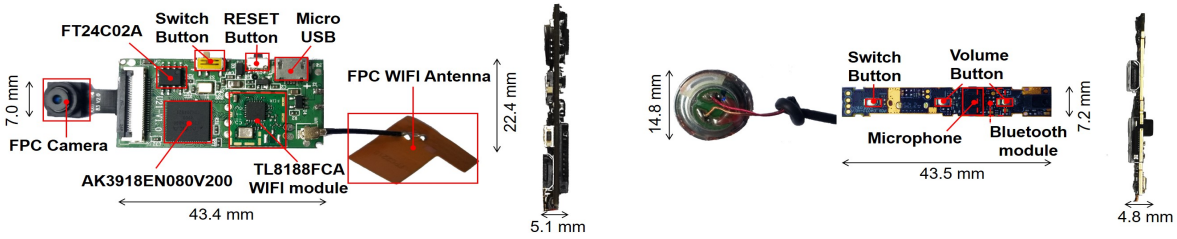
Efficient and intuitive speech guidance: All of the participants mentioned that the guidance should be efficient and intuitive. At the same time, we found that they preferred intuitive speech guidance because of its efficiency and intuition (mentioned 7 times). Only P3 mentioned that he wanted the system to use a vibration prompt, because he thought the sound might cover up the external sound. However, we designed external earphones to prevent the blocking of ambient sound to solve this problem.

3.3.3 Design Requirements Proposed by BLV Users. In addition to functions, to create a complete system for BLV users, we also need to consider the design elements closely related to their daily life, such as form, practicability, and convenience. What system design elements do BLV users care about? We summed up constructive guidelines. We found that usefulness, reliability, and wearability were mentioned by all 8 participants. A discreet shape was also an important factor; it was mentioned 7 times. Then, it should be readily available and easy to operate (mentioned 6 times). Privacy protection was mentioned 4 times. Two participants (P2 and P4) mentioned that it should be a low-cost system.

Usefulness and reliability: All of the participants mentioned that usefulness and reliability are important because they determine whether the system can work normally. They all hoped that the recognition would be as accurate as possible, which could guide them to find target objects in a short time.

Keeping both hands free: The users believed that keeping their hands free would not affect their use of touch and doing other things. During the observations after the interviews, we also found that they always habitually reached out to explore the objects or doors and windows around them. Therefore, the system should have a wearable form that does not affect hand activities. On the contrary, handheld devices are impractical. Ready availability was also an important element. The system should be unrestricted and comfortable to wear so that it can be worn for a long time during the day and used at any time.

Esthetic and inconspicuous appearance: The results of the interview show that users attached great importance to the appearance and form of the device; these factors may even be decisive with respect to whether they were willing to use the device. Four participants (P1, P3, P5, P8) said that they did not want the device to be bulky or conspicuous, which might make their disability more visible. Three participants (P2, P6, P7) explained that they were eager to be treated equally by people with normal vision. Two participants said that a conspicuous appearance might attract unnecessary attention and affect their social interactions and work (P3, P4).



(a) Camera PCB. (b) Earphone PCB.

Fig. 2. Camera PCB and earphones PCB.

4 BOTTOM ELECTRONIC DESIGN AND STRUCTURAL DESIGN

The purpose of this study was to complete the electronic design and structural design of the system based on the functional and design requirements described in Section 3. The electronic design determines the underlying hardware architecture according to the functional requirements and technical conditions of the system. Convenience of the structural design also needs to be considered. The structural design should consider esthetics, comfort, and other design requirements, as well as the layout of the hardware. In short, the electronic design and structural design are interrelated. Therefore, during this process, five researchers conducted several brainstorming sessions and discussions, and structured interviews were conducted with 8 interviewees (P1–P8 in Table 3).

4.1 Hardware Selection and Composition

The choice and coordination of hardware is very important for proper functioning and for the user experience. In order to simultaneously meet the requirements of reliability, usefulness, and wearability, we selected the hardware components that follow.

4.1.1 Binocular Camera. To obtain visual information, a camera is indispensable. However, depth is also important guiding information. Existing technologies for obtaining depth information include the binocular camera distance measurement [60], TOF sensing [61], structured light coding [37], and deep learning [53]. To achieve fast response, we chose a binocular camera instead of additional sensors and complex operations, and used binocular parallax and feature matching based methods to calculate the depth information of the target object and the hand. We chose the flexible printed circuit (FPC) large wide-angle lens module, which has a maximum of 1920×1080 pixels. However, we found that a resolution of only 1280×720 pixels was sufficient to accurately identify objects and hands in the reachable range, so we used a resolution of only 1280×720 pixels to reduce computation.

4.1.2 External Earphones. At present, most ordinary headphones on the market block external sound. To not block the ear canal, a bone conduction headphone to transmit sound through tympanic membrane vibration can be used. However, its sound quality has been criticized (P3, P4, P6, P7). We devised a comprehensive solution that applies the form of an ear-hung headphone and changes the position of the speaker to the front of the ear, but does not block the ear canal (Figure 3). We call this an external earpiece. This keeps the sound not only from being distorted, but also from being heard by people nearby. The 5-point Likert scale showed that BLV users approved our new design: The external headphone was the most popular ($M = 4.3$, $SD = 0.4$), while the bone conduction headphone scored slightly lower ($M = 3.8$, $SD = 1.1$).

4.1.3 Camera PCB and Earphone PCB. As shown in Figure 2, in addition to the FPC camera module, there is an FT24C02A chip and an Anyka AK3918EN080V200 processor (a special multimedia application processor) on the camera printed circuit board (PCB). The FT24C02A chip can connect the camera and the Anyka processor

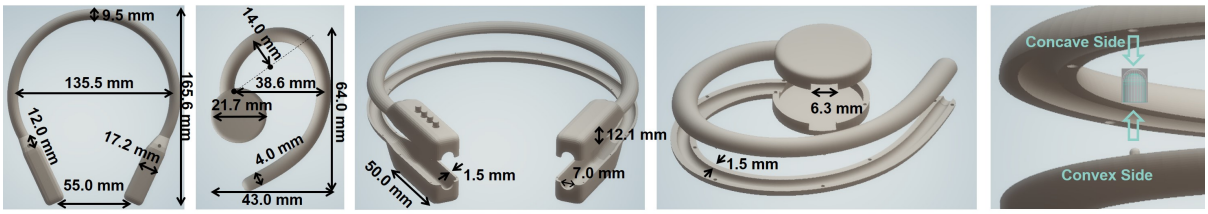


Fig. 3. Dimensions of the shell.

to support data transmission of the I2C protocol between them. The TL8188FCA WiFi module and a 2.4G high-gain FPC WiFi antenna are also embedded in the camera PCB. The TL8188FCA WiFi module complies with the 802.11b/g/n wireless protocol, and the wireless transmission rate can reach up to 150 Mbps. We used the Real-Time Streaming Protocol (RTSP) to transmit video to the fog server at a rate of 25 frames per second (FPS) to ensure a high speed of data processing and low energy consumption of the mobile headphone terminal. As shown in Figure 2, on the earphone PCB, a microphone and Bluetooth communication module are embedded, which are connected with the speaker with wires. The Bluetooth communication module is convenient for information transmission with the fog server and the cloud.

4.1.4 Fog Server and Cloud Service. We used an NVIDIA GeForce RTX 3090 graphics card as the fog server and CUDA to run a large number of GPU cores in parallel, which can accelerate convolution operations in object recognition. The fog server communicates with the headphone terminal in the LAN through WiFi and Bluetooth. It is mainly used for video processing and speech synthesis with frequent calls and high demand for short delay and computing resources. However, the call frequency of speech recognition is low, and the delay requirement is not too high. Therefore, we used the iFLYTEK [9] speech recognition API cloud service to support them.

4.1.5 Battery and Other Interfaces. We used two 500-mAh long-strip-shaped lithium polymer batteries (3.7 V) to power the camera PCB and earphone PCB. Considering the form of the headphone, the long strip shape is selected because it is easy to deploy. The camera PCB also provides a micro USB charging interface and indicator light for easy debugging.

4.2 Views on the Form of Neckband Headphone

We chose the form of the neckband headphone based on the following considerations.

4.2.1 Expandability of the Headphone. A headphone is a wearable device needed by BLV users in their daily life. Originally it was used to play music and as a telephone intercom. Some BLV users do not have the habit of wearing glasses. However, it is easy to accept the extension of an existing headphone when additional functions are added.

4.2.2 Proper Camera Position and Orientation. As most target objects are located on the table or the floor, the camera should be able to capture target objects on the table and the floor. Existing wearable devices are mainly worn on the eyes [54, 84], ears [23, 27, 64], neck [83], shoulders [46], chest [52], wrists [29], hands [45], fingers [22], and legs [16]. We found that the camera installed at the end of the neckband headphone can naturally aim at the table and floor in front of the torso (the angle with the horizontal plane is about 25°–55°) without bowing the head (glasses) or straightening the arm (wrist strap and gloves); besides, it ensures that objects are not blocked.

4.2.3 Suitable for Long-term Use. Compared with glasses resting on the nose bridge and the ears, neckband headphones rest on the whole neck, which reduces the pressure. The headphone can be comfortably hung around the neck and used at any time. It is suitable for long-term use at any time of the day.

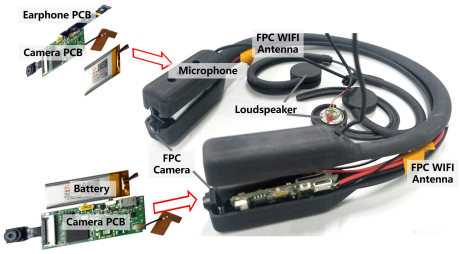


Fig. 4. Hardware layout details.

4.2.4 Enough Space to Maintain an Unobtrusive Shape. In addition, the neckband can provide enough space to accommodate the new camera PCB and wires without a significantly appearance change, and it can also store a large enough battery to support long-term use. On the contrary, the glasses often studied easily attract others' attention because they make it difficult to accommodate more hardware [6, 28, 32, 54, 59, 78], which can be considered as an obvious drawback [65, 79].

4.3 Shell Design and Hardware Layout

4.3.1 Shell Design. We modeled the headphone shell using 3D Studio Max, including a collar and two external headphones. We used acrylonitrile-butadiene-styrene (ABS) material to print the shell through the stereo lithography appearance (SLA) process. The ABS material provides enough stiffness to ensure that the baseline distance between the two cameras does not change, which can provide stable and accurate distance measurement. We split the collar shell into two parts that can be opened to accommodate the hardware, and printed concave and convex edges to facilitate the alignment of the two parts of the case. Two spaces are formed at the two ends of the collar to accommodate the camera PCB, headphone PCB, and battery. The collar has a 7.0-mm diameter circular hollow at the end to mount the camera. We made four buttons: to control the headphone switch, the camera switch, and the volume (up or down). We designed the earphones according to the size of the user's ear and divided the shell into two parts. The circular part can accommodate the speaker, and the rectangle that is hollowed out near the ear allows the sound to be transmitted to the user's ear canal. Their sizes are shown in Figure 3. A shell with a thickness of 1.4 mm serves as a lightweight case. We also designed a variety of collar sizes (covering collar diameters of 135.5–196.5 mm) to meet the requirements and preferences of BLV users.

4.3.2 Hardware Layout. The hardware layout is shown in Figure 4. We placed two camera PCBs at the ends of the collars on the left and right sides. The camera is installed at the end of the collar to ensure that the field of vision is not blocked by the body. In addition, the camera can naturally aim obliquely downward at the table and the floor. Then, the headphone PCB is put on the collar on the right. We use the process of double-layer PCB to make full use of the space of PCB. At the same time, the components are arranged on a long strip-shaped board, which is convenient to be placed in the collar while maintaining the esthetics of the traditional neckband headphone. The camera PCB size is $43.4 \times 22.4 \times 5.1$ mm, while the headphone PCB size is $43.5 \times 7.2 \times 4.8$ mm. The size of the battery is $46.5 \times 17.0 \times 3.1$ mm. The two batteries are placed on the left and right sides to maintain the balance of the collar. The four buttons are arranged on the right side of the end of the collar for easy operation.

5 TOP-LEVEL SOFTWARE DESIGN

5.1 System Interaction Process

We brainstormed with BLV users (P1–P8 in Table 3) on the following topics: (i) Analyzing and modeling the general process of reaching objects. (ii) Achieving efficient and natural interaction. (iii) The actual situations that

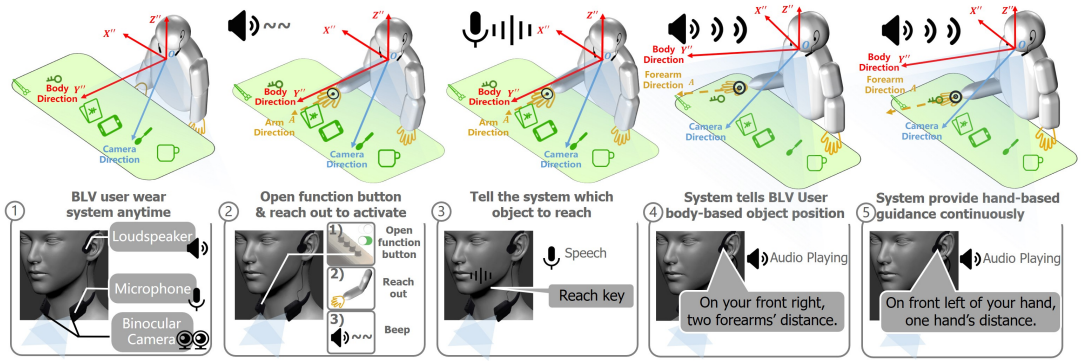


Fig. 5. System interaction process.

may be encountered and solutions to encountered challenges. After brainstorming, we determined a convenient interaction process. Vague memory, too small objects, and fallen objects were several causes of difficulties mentioned. Target objects are often located on a table or on the floor. Therefore, the interactive process is as follows: (1) BLV users wear the system. The system is off or in normal headphone mode. (2) When needed, they press the function button and extend their hand to start. When the hand is detected, a beep is heard, indicating that the system is active. (3) Speech is used to tell the system the target object to reach. (4) The system checks whether the speech contains the target word entered in the database, and then tells the BLV user the position of the target object relative to the torso. (5) When the BLV user's hand is within 35 cm of the target, the system continues to provide speech guidance relative to the hand. The system interaction process is shown in Figure 5. Note that the target object may not be within the field of vision. The interaction of the two situations is as follows: (1) When the target is not detected in the historical frame, the system sends the voice feedback "Please turn your body to continue searching." (2) If the target already appears in the history frame, the short voice feedback "loss" is sent. We found that in some target lost frames, BLV users could accurately locate the target position through historical voice guidance and the feeling of body rotation, and quickly correct the body orientation to the target. This kind of rapid on-line error correction mechanism based on proprioception of BLV users has been reported by some neuroscience studies [36].

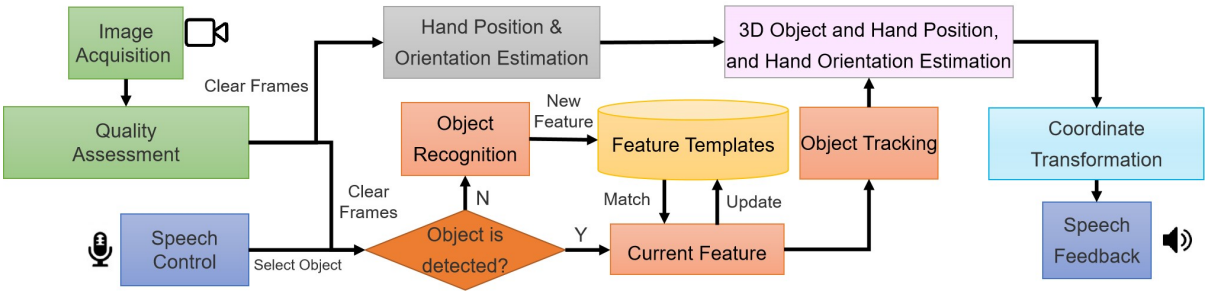


Fig. 6. System workflow.

5.2 System Workflow

At the software level, we designed a series of algorithms based on the interaction process of LiSee and considering the technical level. The system workflow is shown in Figure 6. LiSee first selects clear frames, and then carries out object recognition according to the target object input by BLV users. After the target object is recognized, the feature point template of the target object is obtained. LiSee tracks the target object by matching the feature points of the template in the subsequent frames to reduce calculation. Meanwhile, LiSee estimates the hand pose of BLV users through a deep neural network. Next, LiSee combines depth information to estimate the 3D position of objects and the 3D hand pose. Finally, LiSee provides torso-based and hand-based guidance after coordinate transformation. It is worth mentioning that our series of algorithms follows three design principles: (1) low complexity of the algorithm, (2) independent call of the algorithm, and (3) utilization of intermediate operation results to improve the running speed of LiSee. In the following, we elaborate our design.

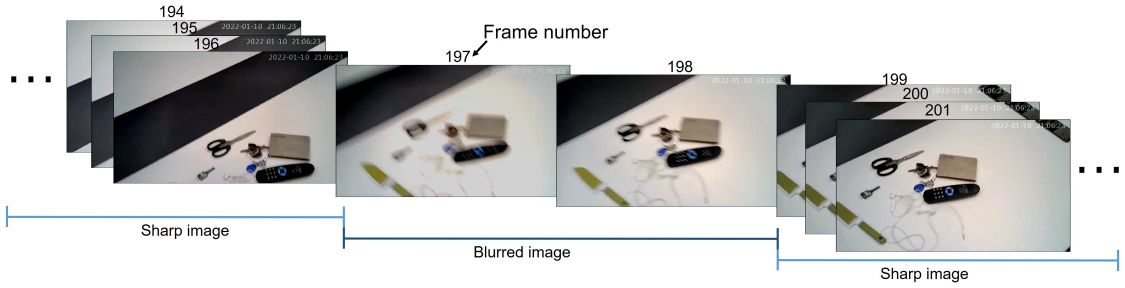


Fig. 7. Continuous frames of images captured by right camera.

5.3 Automatic Clear Frame Selection

We proposed a motion blur detection algorithm based on **the lowest directional high-frequency energy** (the sum of the squared directional derivatives of image) [30]. This index can detect the blurred region effectively without performing the point spread function and deblurring algorithm simultaneously and has high robustness and efficiency [50, 66]. Specifically, the images are divided into 30×30 of 42×24 blocks (block size 30×30 is the experience of [30]), which is recorded as $b_{1,1}, b_{1,2}, ...b_{42,24}$. The derivative along the direction of each pixel can be written as:

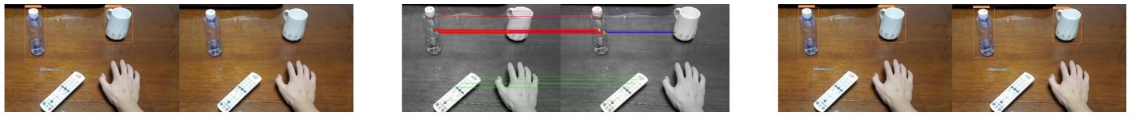
$$\Delta f(x,y)_k = \begin{bmatrix} f'_x & f'_y \end{bmatrix} \begin{bmatrix} \cos(k) \\ sin(k) \end{bmatrix} \tag{1}$$

where $f'_x$and f'_y are the derivatives in the horizontal and vertical directions, respectively. Then, the sum of the squared directional derivatives of the pixel blocks $b_{i,j}$ can be written as:

$$S_{i,j}(k) = \sum_{x=1}^{30}\sum_{y=1}^{30}\left(\begin{bmatrix} f'_x & f'_y \end{bmatrix}\begin{bmatrix} \cos(k) \\ \sin(k) \end{bmatrix}\right)^2 = \begin{bmatrix} \cos(k) \\ \sin(k) \end{bmatrix}^T \left\{\sum_{x=1}^{30}\sum_{y=1}^{30}\begin{bmatrix} f'_x f'_x & f'_x f'_y \\ f'_x f'_y & f'_y f'_y \end{bmatrix}\right\}\begin{bmatrix} \cos(k) \\ \sin(k) \end{bmatrix} \tag{2}$$

The minimum value of $S_{i,j}(k)$ should meet the condition $\frac{d}{dk}S_{i,j}(k) = 0$. We record the minimum value of $S_{i,j}(k)$ as $S_{i,j}(k_{min})$, where k_{min} represents the direction in which the sum of the squared directional derivatives is the smallest, representing the motion direction. However, in the direction perpendicular to the motion direction, the sum of the squared directional derivatives is the largest. We record this direction as k_{max} (i.e., $k_{min} + \frac{\pi}{2}$) and the sum of the squared directional derivatives in this direction as $S_{i,j}(k_{max})$ (i.e., $S_{i,j}(k_{min} + \frac{\pi}{2})$).

As shown in Figure 7, it is obvious that among blocks with objects, the texture and contour details of objects have the most significant stretch size in the motion direction. The ability to retain high frequency energy is strongest in the direction perpendicular to the motion direction and the sum of the squared directional derivatives



(a) Recognition result of previous frame (left) & current frame (right)

(b) ORB feature points template matching

(c) Tracking result of current frame image (right)

Fig. 8. 2D object recognition and tracking.

is greatest [30]. However, for desktop regions of the same color, whether in motion or non-motion, small sums of the squared directional derivatives are maintained in both k_{min} and k_{max}. Therefore, detecting the table and floor blocks has limitations in judging the motion blur. When the objects are not blurry, object blocks maintain a large value in the sums of squares of both k_{min} and k_{max} directions. In order to ensure that the judgment is robust to the image content, we exclude the table and floor blocks with the same color, extract only the textured blocks representing the objects, and calculate the ratio of the number of blurry blocks to the number of clear blocks. According to our experience, we consider the blocks with $S(k_{min}) < 100 \ \ \& \ \ S(k_{max}) > 130$ as blurry blocks, and record the number of blurry blocks as Num_b. On the other hand, we consider the blocks with $S(k_{min}) > 130 \ \ \& \ \ S(k_{max}) > 130$ as non blurry blocks, and record the number of non blurry blocks as Num_{nb}. We empirically consider images that $\frac{Num_b}{Num_{nb}} > 1.2$ as blurred. Note that our judgment is conservative because the motion blur caused by camera movement makes all the object blocks in the entire image blurred.

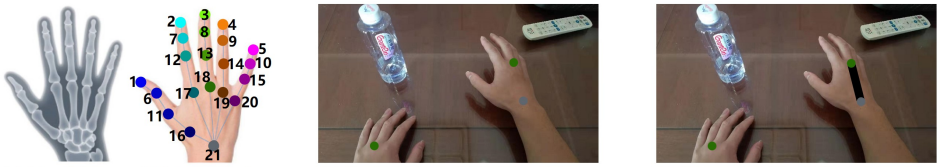
5.4 2D Object Recognition and Tracking

5.4.1 Object Recognition. We chose Mobilenet-SSD [40] as the object recognition network, which is a single-stage multi-scale object detector. Because SSD is single-stage, it is faster than secondary object recognizers (such as Fast R-CNN [34]). Mobilenet's mechanism also further improves the recognition speed. Mobilenet-SSD not only meets the real-time requirements, but its multi-scale characteristics can also detect small enough objects, such as keys. Since there are no datasets for reaching target object task, we retrained a target detector that can distinguish 10 common objects. We collected pictures under different lighting conditions, different table colors or floor colors, and different numbers of objects in the picture to cover the complex environment in actual use. The training process is explained in Section 6.1.1. Note that we only recognize the target object in the right image to reduce the computational workload. After object recognition, we obtain the 2D position of the target object (the bounding box of the target object) and the target object recognition probability (the output of Mobilenet-SSD). Only target objects with a recognition probability greater than 0.6 are considered to be correctly identified. The threshold is set to reduce false recognition.

5.4.2 Object Tracking. If the object bounding box has been recognized in the previous frame, then, for subsequent frames, we do not perform object recognition with relatively large calculations. We can extract the features in the object bounding box of the previous frame and generate the feature point template. Obviously, we only need to match feature points for each subsequent frame to quickly track the object in real time. Specifically, our algorithm is divided into the following steps (Figure 6):

(1) *Judge whether the object is locked*: Judge whether the target object specified by BLV users has been correctly identified in the previous frame. If so, go to step (2). If the target object is not recognized in the historical frame, LiSee continues object recognition. If the feature template of the target object has been obtained from the previous frame, go to step (3).

(2) *Obtain the feature point template of the object*: First, we convert the image into a gray image. Next, we use the ORB feature point detection algorithm [67] to detect only the feature points in the bounding box of the previous frame. The reason for choosing ORB is that it is faster and more robust than SIFT [57] and SURF [20]. We use the obtained feature points as the object feature point template. Note that we do not directly use the



(a) The skeleton and joints distribution of the right hand

(b) The 18th joint and 21st joint of the hands are detected

(c) the metacarpal bone that connects the 21st joint to the 18th joint is connected

Fig. 9. 2D hand position and orientation tracking.

pixels in the entire bounding box as a template because the object may translate, rotate, and scale, thus resulting in the failure of pixel-based matching and a slow speed of matching all pixels.

(3) *Matching and updating templates*: It is assumed that the feature point template of the object has been obtained in the previous frame image. Because the change of object displacement between two adjacent clear frames is very small, we only detect ORB feature points in the object neighborhood of the current frame. Specifically, through experiments, we find that the feature point displacement of the object in the adjacent frame does not exceed 30 pixels. Therefore, assuming that the pixels in the bounding box of the previous frame are (u_{px}, v_{py}), we only detect the feature points in the neighborhood $([u_{px} - 30, u_{px} + 30), (v_{py} - 30, v_{py} + 30)])$ of the current frame. Then, for each pixel, we only use the feature point matching based on the minimum Hamming distance for the neighborhood. We use the cross matching algorithm to verify whether the matching from the current frame to the previous frame is correct. Cross-matching refers to matching the feature points in the previous frame with the feature points in the current frame. A match is considered correct only if the feature points of the previous frame and the current frame match each other. After correct matching, all feature points $p'(x', y')$ of the object have the same affine transformation relative to the feature points $p(x, y)$ of the previous frame, that is, the combination of rotation, scaling and translation. We select the three feature points with the highest confidence to calculate an affine matrix, which represents the transformation of the object of the current frame relative to the previous frame. In this way, we can obtain a new bounding box according to the affine matrix and the bounding box of the previous frame image. At the same time, all feature points outside the bounding box are deleted, and the feature points in the bounding box are regarded as the new template of the current frame. In order to ensure that the object is tracked correctly, when the number of feature points is less than 10, we re-enter the object recognition stage. The general process is shown in Figure 8.

5.5 2D Hand Pose Tracking

Figure 9(a) shows the joint distribution of the hand. Our goal is to obtain the position and direction of the hand. We found that although the hand has 21 joints, we can obtain the position and orientation of the hand according to only the 18th and 21st joints (Figure 9(b)). Specifically, we use the network model of SRHandNet [76] for reference to recognize the posture of the hand. We use the model trained by [76] to obtain the 18th and 21st joints. We regard the position of the 18th joint as the position of the hand, and we regard the metacarpal bone that connects the 21st joint to the 18th joint as the direction of the hand (Figure 9(c)). Note that we may detect two hands, but we only focus on the hand closest to the target object.

5.6 3D Position and Orientation Estimation

After the 2D positions of the hand and object are tracked, we calculate the depth to obtain the 3D position of the object and hand, and hand orientation. We adopted an efficient and robust binocular depth measurement method, namely Semi-Global Matching (SGM) [39], to quickly obtain the 3D position of the object and hand and hand orientation. SGM can obtain smooth parallax maps and is more robust to illumination conditions [73].

(a) Binocular images

(b) The depth of object bounding box (bottle), hand joints and metacarpal bone

Fig. 10. Object and hand distance estimation.

5.6.1 Binocular Camera Correction. There are errors in camera technology and assembly, resulting in distortion of camera imaging. Using the pinhole imaging principle, we can correct the distortion by establishing a model based on the position of pixels and the position of the real world. We took images on chessboard paper and used the stereo camera calibrator of MATLAB to calculate the internal and external parameters of the two cameras as well as the translation and rotation between the two cameras. Note that the binocular camera of the system needs to be calibrated only once after the hardware assembly is completed, and BLV users are not required to perform any additional operations.

5.6.2 Estimation of 3D Object and Hand Position and Hand Orientation. Our algorithm flow is shown in Figure 10. Our goal is to measure the position of the target object and hand, and hand orientation, not the depth of the entire image. Since we have obtained the object bounding box, two key joints of the hand and the metacarpal bone between the 18th joint of the hand and the 21st joint of the hand in the right image (Section 5.4 and Section 5.5), we only calculate the depth of these key areas. Note that the joints and metacarpal bone of the hand are sets of pixels rather than individual pixels. Specifically, assuming that the pixels of the object bounding box, hand joints and metacarpal bone of the image on the right are (u_{rx}, v_{ry}), we only conduct parallax search and depth calculation for the neighborhood on the left image ($[u_{rx} - 96, u_{rx}]$, $[v_{ry} - 2, v_{ry} + 2]$). This not only reduces computation, but also reduces matching errors. Then, according to the pinhole camera model, we obtain the depth of the object, the 18th joint of the hand and the metacarpal bone through the triangulation principle:

$$Y = \frac{bf}{d} \tag{3}$$

where b is the baseline length (55.0 mm), f is the focal length of the camera (2.8 mm), and d is the difference between the abscissa of the key areas in the left and right images. The average depth of each pixel in the object bounding box and the 18th joint of the hand is taken as the depth of the object and hand. The centers of gravity of the object bounding box and the 18th joint of the hand are regarded as the 3D positions of the object and hand, which we denote as (X_o, Y_o, Z_o) and (X_h, Y_h, Z_h), respectively. The mean value of the depth change along the metacarpal bone is taken as the hand orientation and recorded as $\vec{A}$.

5.7 Coordinate Transformation to Align the Torso

Owing to the novel and unrestrained form, we need to face the following challenges before providing guidance. The view of the camera on the collar naturally faces down toward the table or floor. This is different from some traditional glasses-shaped auxiliary devices whose camera orientation is consistent with the torso direction. However, wearing the neckband headphone for a long time may cause the camera to shift left and right. However, we must provide guidance relative to the torso. We designed a method that requires little user effort. We require BLV users to extend their forearms after opening the function button to make it consistent with the torso. This is a simple action to activate the system function. We utilized the hand orientation detection result $\vec{A}$ of Section 5.6.2 as the torso orientation. Next, as shown in Figure 11, we first calculated the angle θ between the XOY plane and

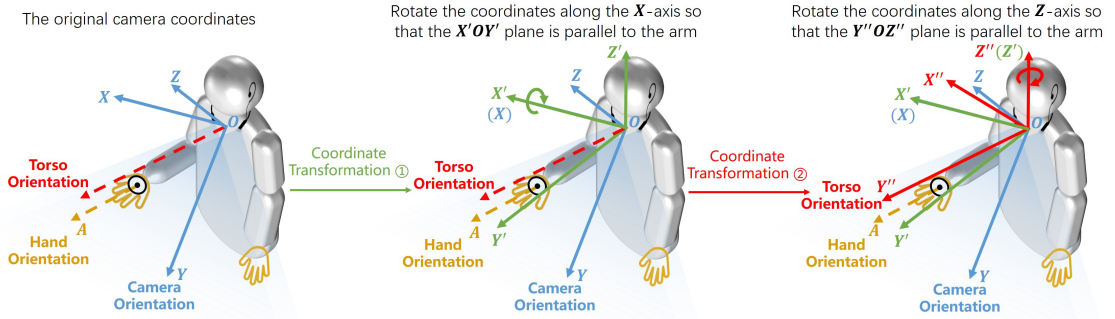


Fig. 11. Coordinate transformation to align with the torso.

the torso orientation, $\theta = arcsin(cos < \vec{A}, \vec{Z} >)$, and then rotated the coordinates along the X axis to complete the first transformation $R_x(\theta)$ so that the $X'OY'$ plane was parallel to the torso orientation. Then, we calculated the angle ϕ between the $Y'OZ'$ plane and torso orientation, $\phi = arcsin(cos < \vec{A}, \vec{X'} >)$, and rotated along the Z axis to complete the second transformation $R_z(\phi)$ so that the $Y''OZ''$ plane was parallel to the torso orientation. In this way, the Y'' axis coordinates coincide with the torso orientation, allowing us to provide torso-based guidance. The complete transformation formula is:

$$\begin{bmatrix} X'' \\ Y'' \\ Z'' \end{bmatrix} = R_Z(\phi)R_X(\theta) \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, R_Z(\phi) = \begin{bmatrix} cos(\phi) & sin(\phi) & 0 \\ -sin(\phi) & cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix}, R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & sin(\theta) \\ 0 & -sin(\theta) & cos(\theta) \end{bmatrix} \quad (4)$$

5.8 Guidance Scheme

From the observation of BLV users reaching target objects, we found that they could only randomly explore the table or floor because they did not know the specific location of the object. Many times, the BLV users started to explore from the opposite position, which consumed a large amount of time. We gained key insights from the BLV users. We used speech guidance and divided the guidance into two stages: coarse-grained guidance and fine-grained guidance. The purpose of coarse-grained guidance is to tell users the approximate position of the object based on the torso (step (4) in Section 5.1) so as to help them quickly adjust their torso orientation or extend their hands corresponding to the orientation. This stage is continuously adjusted until the distance between the hand and the object is $\leq$ 35 cm. On this basis, fine-grained guidance provides hand-based guidance to finely adjust the position of the hand (step (5) in Section 5.1). In further interviews, we found that owing to everyone's expertise and experience, their understanding and preferences for direction and distance were different. Habits and preferences can be overwhelming and difficult to change. Therefore, in order to help users master them quickly, we provide the choices in Table 1. In particular, Figure 12 shows the corresponding areas in the direction guidance scheme. They can choose different combinations based on their preferences and expertise.

Table 1. Guidance schemes that can be selected by BLV users.

	Coarse-grained Guidance	Fine-grained Guidance
Direction	**$Crs_{Dir}1$** : Left (Right) / Front **$Crs_{Dir}2$** : Front of left (right) hand / Left (Right) / Left (Right) front **$Crs_{Dir}3$** : Clock direction (9 / 10 / 11 / 12 / 1 / 2 / 3 o'clock)	**$Fine_{Dir}1$** : Front (Rear) / Left (Right) **$Fine_{Dir}2$** : Front (rear) / Left (Right) / Left (Right) front / Left (Right) rear **$Fine_{Dir}1$** : Clock direction (1-12 o'clock)
Distance unit	**$Crs_{Dis}1$** : cm **$Crs_{Dis}2$** : Length of a forearm (35 cm) **$Crs_{Dis}3$** : Length of a hand (16 cm) **$Crs_{Dis}4$** : Not need	**$Fine_{Dis}1$** : cm **$Fine_{Dis}2$** : Length of a hand (16 cm) **$Fine_{Dis}3$** : Not need

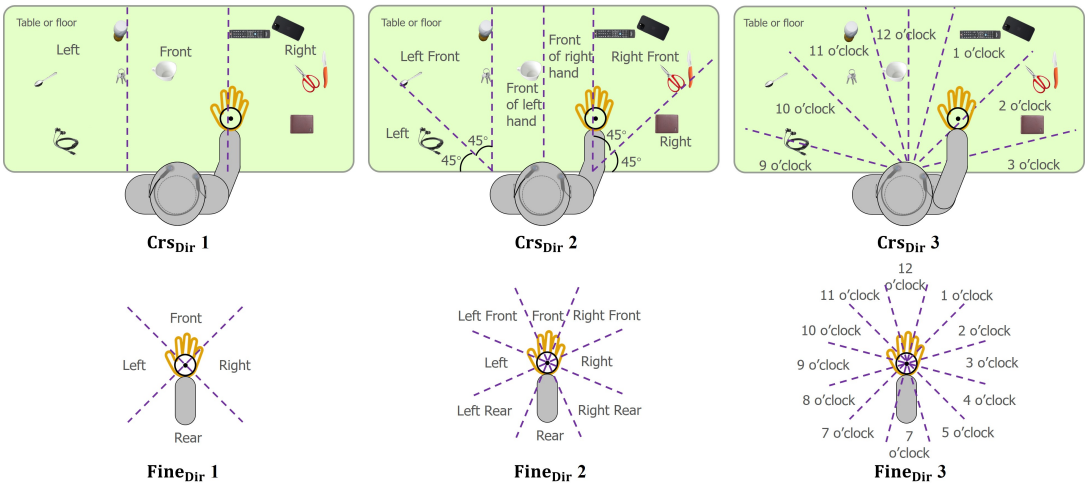


Fig. 12. Detailed corresponding areas of direction guidance schemes (Table 1).

(a) Gray table under weak natural light (b) Yellow table under natural light (c) Yellow table under fluorescent light (d) Brown table under weak natural light + light on

Fig. 13. Some different table color samples under different ambient lighting.

6 EVALUATION

We first evaluated the technical specifications of the system. Second, to study the effect of LiSee, we designed some laboratory tasks and quantified the performance of BLV participants. Finally, to verify whether LiSee can provide BLV users with all-day assistance outside the lab, we studied BLV participants' usage at home for 10 consecutive days through video analysis and user feedback.

6.1 System Technical Evaluation

We evaluated the series of proposed algorithms. We described the training process of object recognizer and evaluated the environmental impact of the object recognizer and the impact of sample size. At present, there is no unified evaluation index for clear frame selection. We therefore estimated the accuracy of object recognition, and observed the effect of clear frame selection. We also evaluated the hand pose recognition. Then, we evaluated errors in 3D position and orientation estimation, with the aim of evaluating 2D position errors for object and hand tracking and combined errors for depth calculations. Finally, we evaluated the power consumption and delay of LiSee.

6.1.1 Training of Object Recognizer. Considering that BLV users should be retrained with as few datasets as possible, we used the training method of transfer learning. First, we let Mobilenet-SSD learn large sample data on

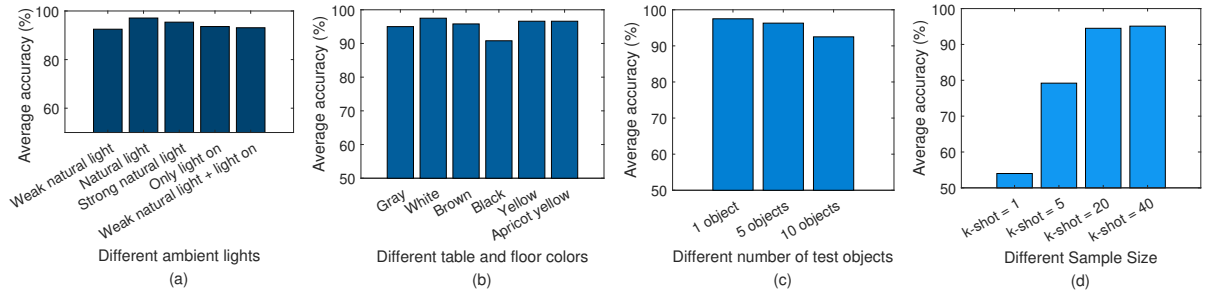


Fig. 14. Average accuracy of the object recognizer under different experimental conditions.

the combined dataset of VOC2007 [1] and VOC2012 [2]. On an NVIDIA GeForce RTX 3090, we used the Pytorch training framework to carry out 40000 training iterations with an initial learning rate of 0.001, and reduced the learning rate by 10 times in 10000 and 20000 iterations. We set the batch size at 32 and used a weight decay of 0.0005 and a momentum of 0.9. After obtaining the network trained on VOC dataset (large sample dataset), we used transfer learning to realize small sample object detection. Intuitively, the lower layer of Mobilenet-SSD has learned to distinguish the characteristics of objects. Specifically, what we do is freeze the final fully connected layer, retain the weight parameters of the feature extractor, and only use the new dataset to train the classifiers (fully connected layer and Softmax) grafted on the feature extractor. Then, the model batch size trained on the small-sample dataset is set to 16, the learning rate is set to 0.0001, and other learning strategies remain unchanged from those on the VOC dataset. In order to cover the use as much as possible, we asked one of the BLV users (P2) to wear LiSee (LiSee's camera was turned on and the clear frame selection algorithm was turned on). P2 initially stood 20 cm from the table and 30 cm in front of objects on the floor and only used touch to explore the target object. Photos were collected under different ambient lights (weak natural light, natural light, strong natural light, only light on, and weak natural light + light on), different table and floor colors (gray, white, brown, black, yellow, and apricot yellow), and different numbers of objects (1–10) in different places in our startup campus (Figure 13 shows some samples). P2 can move and turn her body to approach and reach target objects. The placement of objects is random to simulate the uncertainty of the real world. We collected a total of 6000 photos of 33000 objects (300 environments ($5 \times 6 \times 10$) and 20 photos in each environment). Before our model training, we also adopted a series of data augmentation strategies for small-sample datasets, including random flip, random clipping, deformation scaling, slight blurring, color transformation, and adding random noise. In addition to the above basic data enhancement strategies, we also use Cutmix [82] to process images, which improved the accuracy of the classifier by more than 2 percentage points.

6.1.2 Impact of Different Ambient Lights. Our experimental setup is shown in Figure 16. We let P4 wear LiSee and stand still to take photos. The placement position of 5 objects captured each time is different and random. We tested it in the morning, noon, and evening of the day. Each light condition included 20 photos on the table and the floor. Figure 14(a) shows that the model performs well when taking test images in medium brightness, and natural light is the best. In the case of weak natural light, the decline of image acquisition quality affects the accuracy. If there is additional light with moderate brightness in the case of weak natural light, the accuracy will return to the level of natural light. However, smooth table surfaces reflect light and decrease accuracy.

6.1.3 Impact of Different Image Background Colors. We tested six table or floor background colors under natural light conditions. We let P4 wear LiSee and stand still to take photos. Each background color contained 20 test photos, and the placement of 5 objects in each test was random. As shown in Figure 14(b), the table background

of other colors had little effect on the performance of the object recognizer, except that some objects with a black background were difficult to recognize due to their similar colors. The recognition accuracy with a white table background was the highest (97.5%).

6.1.4 Impact of the Number of Test Objects. In order to explore the influence of the number of objects on the accuracy of object recognizer, we let P4 stand in front of the table and floor (Figure 16) under natural light. Each number of test objects has 20 pictures. When we randomly placed different kinds of recognizable objects in the test area, there might be overlap between objects. As we can see from Figure 14(c) , when there is only one recognizable object in the detection area, the accuracy of our object recognizer is the highest (97.5%). With an increasing number of objects in the detection area, the accuracy of the object recognizer decreases. When there are 10 different objects in the detection area at the same time, the accuracy is the lowest (92.5%). When the number of categories of objects to be detected increases, the overlap of objects will make the object bounding box unstable.

6.1.5 Effect of Sample Size. Under the experimental conditions of natural light and Figure 16, the detection accuracy experiments of 5 objects were carried out on the model. After transferring the new data set of the original model, the model test results of 1, 5, 20 and 40 training samples are shown in Figure 14(d). In a single example study, the average accuracy of the recognizer of blind participants was 54.0%. Among all of the participants, the higher the k, the better the accuracy. We believe that in k-shot learning, the existence of other regions (i.e., not representing objects) in the training data will cause deviation from the solution and reduce the classification performance. We found that sufficient accuracy could be achieved with 20 original training samples (94.5%).

6.1.6 Effectiveness of Clear Frames Selection. According to the setting of collection training set (Section 6.1.1), P4 collected photos under different ambient lights, with different table and floor colors, and with different numbers of objects. We collected 1500 photos of 15000 objects, 138 of which were tested as blurry. The results are shown in Table 2. The average accuracy reaches 96.2%. Precision drops to 92.8% without using clear frame selection techniques. This indicates the usefulness of the clear frame selection technique. LiSee can accurately identify objects that BLV users need to reach.

Table 2. Precision of object recognition. "w/o CFS" means "without using clear frame selection technique"

.

Category	Key	Bottle	Wallet	Cup	Knife	Spoon	Remote	Headphone	Scissors	Phone	Average
Accuracy (%)	95.7	96.6	97.5	95.3	96.6	95.0	97.2	98.3	94.5	95.3	96.2
Accuracy (w/o CFS) (%)	93.3	95.0	91.8	92.5	91.9	93.3	91.5	94.6	93.0	90.6	92.8

6.1.7 Accuracy of Hand Pose Recognition. We performed hand pose recognition on 1362 clear images collected in Section 6.1.6. The True positive rate of hand recognition (two key joints) was 98.5%, indicating that hands were easily recognized. true negative rate is 99.0%, which indicates that there was little misidentification when no hand appeared. The false positive rate was 1.0%, indicating an extremely low probability of identifying other objects as hands. The false negative rate was 1.5%, which indicates that LiSee failed to recognize the shot in a few frames of within a very short time, and LiSee can quickly recognize the shot in later frames. In general, hands can be robustly recognized in different environments.

6.1.8 3D Position and Orientation Estimation Errors. We fixed a high-precision Kinect depth camera [37] (within 4 mm error) on P4's chest to measure the Ground Truth. The calculations of Ground Truth of the position of objects and joints and metacarpal bone orientation were consistent with our calculations (Section 5.6.2). According to the

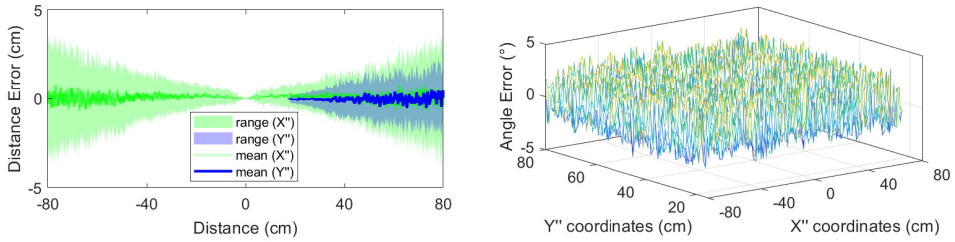


Fig. 15. 3D position estimation errors.

setting of collection training set (Section 6.1.1), P4 collected photos under different ambient lights, with different table and floor colors, and with different number of objects. We collected 1500 clear images and evaluated 8250 objects (we did not eliminate erroneous object recognition because it did not affect this part of the evaluation). Figure 15 shows the 3D estimation errors of object and hand position and hand orientation error (angle error in the $X''OY''$ plane) at different distances. We only calculated the angle error in the $X''OY''$ plane because BLV users reach objects in this plane. The mean value of the position error is near the 0 axis, indicating that there is no calculation error. It can be also seen that as the distance increases, the distance error increases. This is because the parallax decreases as the distance increases, making the calculation more difficult. In summary, position estimation errors within 3.74 cm and hand orientation estimation errors within 4.4° were sufficient for the task of reaching the object. SGM algorithm shows good robustness to table and floor colors and light changes [73].

6.1.9 Power Consumption. As the power consumption limitation of LiSee mainly lies in the earphone end, we used an IT9100 power meter [10] to test the power consumption of the earphone end to see how long the system can be used. We tested the average power consumption of P2 within 65 s of continuously reaching three target objects in front of the table (Figure 16). The camera PCB and earphone PCB consumed 1063.2 mW and 47.9 mW, respectively. Therefore, the earphone terminal consumed 2172.5 mW (2 × 1062.3 mW + 47.9 mW) in total during use. The startup time was about 5 s, and the average power of the camera PCB and headset PCB was 172.8 mW and 11.3 mW, respectively. Therefore, 356.9 mW (2 × 172.8 mW + 11.3 mW) was consumed during startup. Assuming that each startup takes 5 s and the use time is 25 s, the system can be used about 237 ($\frac{3.7V\times500mAh\times3600s/h\times2}{25s\times2172.5mW+5s\times356.9mW}$) times when driven by two 500-mAh batteries (3.7V).

6.1.10 Delay Analysis. Delay measures the real-time performance of the mobile system. Therefore, we tested the delay of each stage according to LiSee's architecture, including the following three parts:

Speech recognition end-to-end delay. This part of delay includes the delay of voice sending to the cloud, iFLYTEK API voice recognition, and sending voice back to the server. We recorded an instruction from one of our researchers (i.e., "reach cup") and obtained a sound waveform with time stamp. On the server side, we obtained the time stamp when the speech transcription was completed with time.time(). We subtracted the time stamp at the end of speech input from the time stamp at the end of speech transcription and found that the delay was about 362.8 ms.

Image processing delay. The original frame rate of image transmission was 25 FPS, so the sum of the delay of image acquisition at the earphone and the delay of WiFi transmission to the server was 40 ms. On the server side, we tested the delay of image processing technology, and the result was 28 ms. Therefore, the total delay of the image processing part was 68.0 ms.

End-to-end delay of voice feedback. We recorded the time stamp and emitted a voice (i.e., "on your front right"), and then captured the sound waveform with time stamp at the speaker of the earphone. This part of delay was about 17.1 ms. In sum, the delay of the three parts added up to 447.9 ms.

6.2 Pilot Study: Laboratory

To form an initial impression of LiSee's impact on BLV users, we first conducted a pilot study in the lab to obtain quantitative results of system usability. We also analyzed the impact of other factors on BLV users' behavior.

6.2.1 Participants. In addition to the participants who participated in the system design (P1–P8), we recruited 4 BLV users (P9–P12) to evaluate more broadly. The demographic information is shown in Table 3. All of the participants said they had no hearing difficulty (this was not tested for confirmation). Our experiment was done in accordance with the Declaration of Helsinki [17], and participants provided informed consent.

Table 3. Detailed demographic information and SUS scores of 12 BLV participants.

Participant	Gender	Age	Vision condition	Diagnosis	Age of vision loss	Education	Occupation	Hearing difficulty	SUS score
P1	F	24	Blind	Glaucoma	0	University	Undergraduate student	No	84
P2	F	36	Low vision Visual acuity=0.1	Retinitis Pigmentosa	16	High school	Beautician	No	86
P3	M	45	Low vision Visual acuity=0.2	Optic Atrophy	12	High school	Blind masseur	No	75
P4	M	33	Blind	Glaucoma	9	University	Accessibility engineer	No	87
P5	F	58	Blind	Cataract	25	Primary school	Unemployed	No	74
P6	M	49	Blind	Birth Blind	0	Primary school	Blind masseur	No	81
P7	M	27	Low vision Visual acuity=0.2	Glaucoma	7	University	Music practitioners	No	87
P8	F	39	Low vision Visual acuity=0.1	Retinitis Pigmentosa	9	Junior high school	Blind masseur	No	80
P9	F	43	Low vision Visual acuity=0.1	Cataract	13	Technical school	Blind masseur	No	79
P10	M	58	Blind	Cataract	14	Primary school	Unemployed	No	87
P11	F	23	Low vision Visual acuity=0.2	Glaucoma	21	High school	Undergraduate student	No	91
P12	M	32	Blind	Birth Blind	0	Primary school	Blind masseur	No	85

6.2.2 Procedure. We conducted standardized experiments on a table and floor in a lab in the startup campus. Participants could choose the appropriate collar size and adjust the volume of their headphone. Then, participants were reintroduced to various guidance schemes and spent 30 minutes trying different combinations of guidance schemes under the setup described in Section 6.2.3. After practice, they chose their preferred guidance schemes to perform the following tasks. Their choices are shown in Table 4.

Table 4. Selection of distance guidance and direction guidance schemes in different stages.

Participants	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Direction scheme in coarse-grained guidance	Crs_{Dir} 2	Crs_{Dir} 1	Crs_{Dir} 2	Crs_{Dir} 2	Crs_{Dir} 2	Crs_{Dir} 2	Crs_{Dir}3	Crs_{Dir} 3	Crs_{Dir} 1	Crs_{Dir}1	Crs_{Dir} 3	Crs_{Dir} 2
Distance scheme in coarse-grained guidance	Crs_{Dis} 2	Crs_{Dis} 2	Crs_{Dis} 1	Crs_{Dis} 1	Crs_{Dis} 2	Crs_{Dis} 1	Crs_{Dis} 3	Crs_{Dis} 4	Crs_{Dis} 2	Crs_{Dis} 1	Crs_{Dis} 2	Crs_{Dis} 2
Direction scheme in fine-grained guidance	$Fine_{Dir}$ 2	$Fine_{Dir}$ 1	$Fine_{Dir}$ 2	$Fine_{Dir}$ 1	$Fine_{Dir}$ 2	$Fine_{Dir}$ 2	$Fine_{Dir}$ 3	$Fine_{Dir}$ 2	$Fine_{Dir}$ 1	$Fine_{Dir}$ 2	$Fine_{Dir}$ 1	$Fine_{Dir}$ 2
Distance scheme in fine-grained guidance	$Fine_{Dis}$ 2	$Fine_{Dis}$ 2	$Fine_{Dis}$ 2	$Fine_{Dis}$ 1	$Fine_{Dis}$ 2	$Fine_{Dis}$ 2	$Fine_{Dis}$ 3	$Fine_{Dis}$ 3	$Fine_{Dis}$ 1	$Fine_{Dis}$ 3	$Fine_{Dis}$ 1	$Fine_{Dis}$ 2

6.2.3 Task. As shown in Figure 16, participants initially stood 20 cm from the table and 30 cm in front of the objects on the floor. Ten objects in Table 2 were placed randomly on an 87 cm × 165 cm table (table height, 110 cm) and a 103 cm × 148 cm floor under natural light. Each target object was treated as a separate task, and this experiment included 10 table and 10 floor tasks. To better understand and compare the impact of LiSee on the users, we tested the users' baseline tasks and the tasks wearing the device.

Baseline tasks: The BLV user searches the target object with the dominant hand without system assistance.

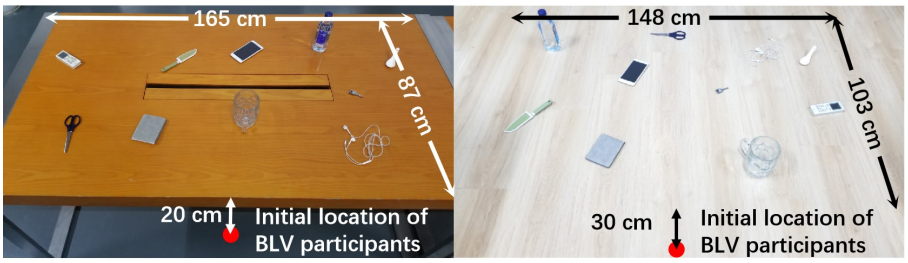


Fig. 16. Setting of laboratory experiment. Left: Table tasks. Right: Floor tasks (natural light).

Device-enabled tasks: Tasks were performed by users using LiSee. We told them not to explore on their own in these tasks so as not to influence the evaluation. In order to eliminate the influence of memory on task order, we rearranged the objects randomly after baseline tasks. We also rearranged the objects for each participant to maximize the randomness of the objects' positions. We told them in a low voice what target object to reach, and then record the time they spent on each task. In both types of experiments, BLV users can move and turn their body to approach and reach target objects. In order to exclude outliers, each task had a time limit of 60 s [72].

6.2.4 Result. **Success Rate and Time Consumption of Reaching.** The overall success rate of all of the participants at baseline was 93.3%, and the overall success rate using LiSee was 96.25%. We analyzed the failure cases later. The average time for each participant to reach the target object after excluding failure cases is shown in Figure 17. In the absence of LiSee, the time they spent was 24.5 s and the standard deviation is 4.5 s. With the help of LiSee, their average time to reach target objects was 14.7 s, their standard deviation was 2.6 s. It can be seen that although the guidance schemes selected by the participants are different, the time they took to reach the target was significantly shorter than the baseline time, which indicates that participants can select the appropriate guidance scheme and quickly reach the target with the assistance of LiSee.

The Hand Track. The trajectories of P3 and P4 reaching the knife on the table with and without LiSee are shown in Figure 18. In the baseline task, the trajectories of the hand resemble random global searches. LiSee allowed participants to locate the object and reach straight for it. The results also show the usefulness of our direction scheme in coarse-grained guidance, especially when the target object direction is known, so that they can reach the target object with the correct hand.

Impact of the Eye Conditions. As shown in Figure 17, we found that low-vision users (especially P2 and P9) spent less time than blind users (especially P1, P6 and P12) in baseline. This situation is easy to understand because low-vision users can use the residual vision to help locate the target object. However, the benefit of residual vision of low-vision users is related to the specific central vision. Therefore, some low-vision users also spent more time, such as P3 and P11. The average time needed by low-vision users was 23.8 s, while the average time needed by blind users was 25.2 s, which is not statistically significant. Moreover, with the help of LiSee, even blind users can reach objects in a very short time, even surpassing low-vision users, such as P1, P4, and P5.

Impact of Table and Floor. As shown in Figure 19, we found that the time to reach target objects on the table (M=13.8, SD=2.4) is shorter than the time to reach target objects on the floor (M = 15.5, SD = 2.8). Because the distance between the camera and the table does not change much when performing tasks on the table, good vision and stability are ensured. However, participants had to squat to reach the target objects on the floor, which shortened the distance between the camera and objects, and narrowed the field of vision, requiring the user to turn their body to reidentify the target objects.

Effect of Object Size. The object size may be a factor affecting the capture time. We averaged the time for all of the participants to reach different objects. As shown in Figure 20, participants spent more time grabbing

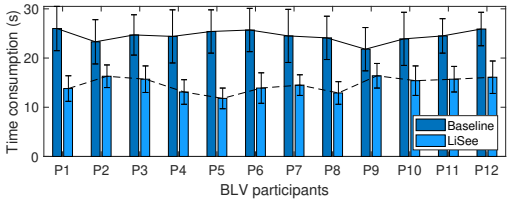


Fig. 17. Time consumption of 12 participants at baseline and with LiSee.

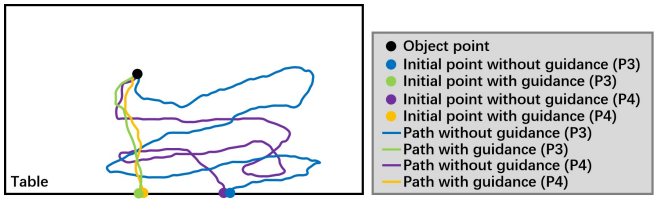


Fig. 18. Trajectories of P3 and P4 reaching the knife on the table with and without LiSee.

smaller objects (e.g., keys, remote controls and spoons) with and without LiSee, which echoed the interview results described in Section 3.3.1 (tiny objects are pain points). Smaller objects easily slip away from the hand in fine-grained guidance. When the participants touch a larger object, they grasp them quickly. Regardless of object size, LiSee helps BLV users reach objects faster.

Effect of Object Positions. We regarded 50 cm as the threshold to define far and near distances. The time needed to reach far and near target objects is shown in Figure 21. As the distance increased, participants spent more time to reach the target objects. Intuitively, participants spent slightly less time on close objects than on far objects. We also noted that the time to reach the objects on the right is about 1.4 s shorter than the time to reach the objects on the left. This may be because most participants are right-handed.

Video Analysis. In the above evaluation, we also observed participants' behavior. We found that P9's perception of forearm length was not very accurate in the coarse-grained guidance stage. He extended his hand too far in the first four tasks, and sometimes his forearm covered the target objects. He reidentified the objects by moving his hand, which took a long time. As P9 became more familiar with LiSee, this occurred less. We also found that P3, P7, and P11 habitually leaned forward at the beginning of the task, sometimes resulting in the target objects not being recognized because they were out of camera view. P7 explained to us that he was a glaucoma patient who could see in a very small area in the center of the field of vision, which is why he leaned forward and used residual vision to feel the target. At the prompt of LiSee, he quickly adjusted the camera's field of view.

Failure Cases Analysis. We found that the main reasons for failure in the baseline task were that the participants did not have a fixed search order and the searches were too random. In the case of failure, they forgot which locations they had searched and they did not know which locations had not yet been searched. When using LiSee, the following technical failures were observed: 1) Object recognition errors. In some cases, the object color was similar to the table color (knife), the tiny target object was blocked (key and spoon), or the object was on the floor and the target object was too tiny (key and spoon). Sometimes, the object recognizer recognized other objects as target objects (e.g., a cup was recognized as a bottle). 2) Hand recognition errors. LiSee sometimes mistakenly recognized other pixels as hand joints. Sometimes the hand orientation estimation was wrong due to wrong judgment of joints. Note that, there was no case where the hand was too small to be recognized, because the distance between the hand and the camera was less than 90 cm. 3) Failure to estimate the 3D position. Depth calculation showed good robustness in this study. In four of the failure cases, the main reason was the error of object and hand position recognition. The main reason for human failure was that the object was not in the field of view of the camera.

6.3 Field Study: Home

The pilot study in the lab showed the significant usefulness of LiSee. Next, we wanted to see how BLV users actually use our system at home, where the light environment and table or floor situations are more complex.

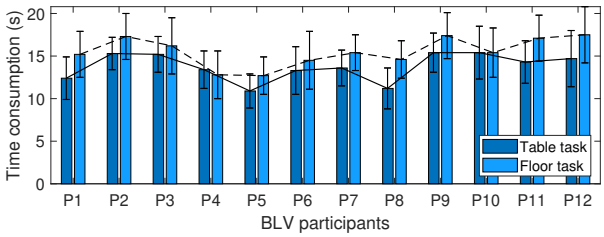


Fig. 19. Differences in time consumption between table and floor tasks of 12 participants.

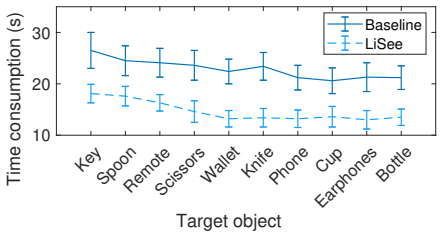


Fig. 20. Impact of size of target objects on time consumption.

6.3.1 Procedure. Because P2, P5, and P6 were unable to participate in the experiment at home for personal reasons, we deployed the whole system in the homes of the remaining 9 participates at different times. We conducted a 10-day study, during which users could use LiSee according to their own needs. At the beginning, the power of LiSee was 100%. With their consent, we inserted a micro SD card into the main board of the camera to record the video shot by the camera of LiSee. Finally, each participant completed a standardized System Usability Scale (SUS) [26] after 10 days of use, as well as a semi-structured interview to gather qualitative insights into LiSee. P2, P5, and P6 also participated in the SUS scoring only based on their experience in the laboratory. We told participants how to turn LiSee on and off and encouraged them to use LiSee in a variety of environments.

6.3.2 Video Analysis. We collected a total of about 3 hours of usage video. We analyzed the use environments, user behaviors and failure cases. We made the following observations. Almost all users used LiSee every day, and the success rate was 94.1%. However, as shown in Figure 22, they spent more time than in the laboratory (after excluding failure cases, M = 21.5, SD = 5.0), which may be because home scenes are more complex. We found that the places where the users reached the objects were mainly the table of the dining room and the table of the living room. Sometimes they picked up objects on the floor. Ambient light conditions included weak natural light, natural light, strong natural light, and light on. The colors of the table and floor include white, gray, brown, pink and yellow. In addition to 10 registered objects, their table also included other things, such as brushes, plastic bags, and masks.

In case of weak light or reflection, LiSee cannot recognize the target objects or mistakenly recognizes other objects. P5's desk is backed by light, which makes the field of vision of LiSee's camera very dark and renders LiSee unable to recognize the key. P9's smooth desktop reflects light, resulting in the wrong recognition of objects in reflective places. We also found that the target object was lost in some cases, especially when the target object was on the floor, but participants could rerecognize the object after body adjustment. Sometimes participants failed because the target object was not on the current table, but they might find the target object on another table later. Some users failed to recognize categories that we have not preregistered. In fact, we found that most participants did not have much on their desktop, but included multiple objects of the same kind. For example, most of them had multiple cups. In this case, LiSee feeds back the nearest object. P11's desk was messy, which led LiSee to treat her wallet as a mobile phone. In a P6 failure case, LiSee also mistakenly identified the edge of the table as the 18th hand joint, which may be because the P6's table has a pink color similar to the hand. There were some differences in time consumption between users. For example, P4 and P9 took nearly 7 s less to complete tasks than P11. This is not only related to their environment, but also to their expertise.

6.3.3 Subjective Rating. We used the SUS score [26] to quantify the usability of LiSee. Participants were asked to rate LiSee on a five-point scale from "strongly agree" to "strongly disagree" on 10 usability questions. Their ratings are shown in Table 3. We found that the average SUS score of LiSee was 83 and the standard deviation

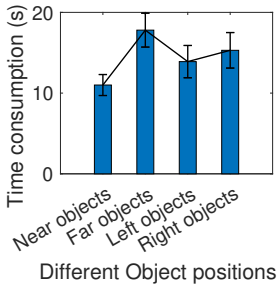


Fig. 21. Impact of object positions on time consumption.

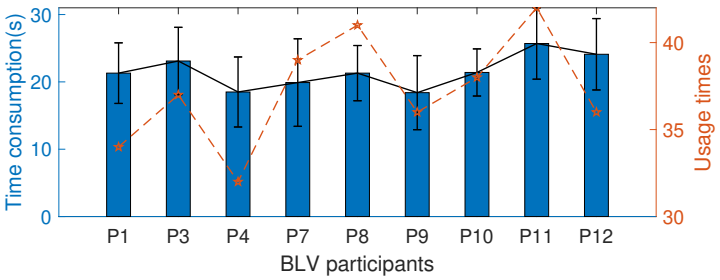


Fig. 22. Usage times and time consumption of 9 participants at home.

was 5. This indicates that the usability of our system is good [19]. SUS question 1 (I think that I would like to use this system frequently) scored as high as 92 (SD = 3), indicating that participants were willing to use LiSee frequently in their daily lives.

6.3.4 Qualitative Feedback. Most participants affirmed the usefulness and reliability of LiSee. For example, P1 stated: "*LiSee enables me to know the location of target objects and helps me to reach the objects. So I don't have to search with my hands. I'm glad it saved me a lot of time.*" P3 stated: "*Speech guides me to my target objects quickly, so I don't have to search by touch or rely on my family.*" P9 stated: "*I can reach the target object accurately according to the speech guidance, so I don't have to spend energy to remember the location of the objects, which is convenient.*" Participants were also excited about the form of headphone. P7 stated: "*I wear headphone almost every day and it's incredible that I can not only listen to music but also reach target objects around me.*" P11 stated: "*I like to wear it, and I wear it all day long and I put the headphone on my ears when I want to reach an object. Sometimes I even forget the headphone is there.*" However, they also offer some suggestions for improvements, such as increasing the types of recognition and increasing the camera's field of view.

7 DISCUSSION

7.1 The Benefits of the Form of the Neckband Headphone

The main purpose of the present study is to design a wearable device suitable for all-day use. The neckband headphone we proposed is not only more suitable for the task of reaching target objects (Section 4.2), but also has better wearability, comfort, acceptability, camera stability and endurance than the current wearable system proposed for BLV users. We compared LiSee in terms of form design, with smart gloves [55, 72], smart glasses [6, 59], smart bone conduction headphone [23], and mobile phones hanging in front of the chest [49, 52].

7.1.1 Wearability and Comfort. As mentioned earlier, smart gloves are not suitable for all-day use, and BLV users often wear them only when they are required for the task at hand. This frequent wearing and taking off is not really convenient. LiSee is similar to smart glasses and smart bone conduction headphones. After adding the camera PCB, earphone PCB, and battery to ordinary equipment, the increased equipment weight has an impact on the comfort of wearing. However, the force bearing points of LiSee are different from those of smart glasses and smart bone conduction headphones. Smart glasses rest on the nose bridge and the ears, and smart bone conduction earphones rest on two ears. LiSee rests on the whole neck. The low pressure renders the neckband headphone more suitable for all-day use.

7.1.2 Acceptability. According to the interviews (Section 3.3.3), BLV users believe that wearable devices need to be esthetic and unobtrusive. Smart gloves are more conspicuous because they are inconvenient to encapsulate into ordinary gloves. The prominent camera in front of the frame of smart glasses (e.g., Google Glass) is unacceptable for most users [31]. We believe that the simple glasses frame does not have enough space to accommodate the camera and large battery. In contrast, smart bone conduction headphones and LiSee are more acceptable visually because they are closer to the ordinary bone conduction headphone and neckband headphone.

7.1.3 Proper Camera Orientation and Stability. Smart gloves require BLV users to straighten their arms and aim at the target object. The camera of smart glasses often faces directly in front, so BLV users need to lower their head to aim at the table or floor. The camera of the smart bone conduction headphone is easily blocked by the head [23], impeding application of the binocular ranging scheme. Although the mobile phone hanging on the chest performs well with respect to acceptability and wearability, its camera orientation needs to be adjusted manually, and its stability is poor. In contrast, LiSee's camera tilts downward and naturally aims at the table or floor. Moreover, LiSee is relatively fixed to the neck, which can well ensure the stability of the camera.

7.1.4 Endurance. LiSee can be used 237 times on one battery charge. Conservatively, BLV users can use LiSee for two weeks. Wearing LiSee for a long time requires long endurance. On the one hand, the collar form of the neckband headphone we proposed provides enough space for the placement of the batteries. On the other hand, considering the limited capacity of the batteries at the headphone end, we only carried out image acquisition, voice acquisition, voice playback and communication at the mobile end. We offload the calculation of the intensive image processing part with serious power consumption to the server. We suggest considering the endurance when designing wearable devices for BLV users, because BLV users prefer to use them for a long time without frequent charging.

7.2 Technical Discussion

7.2.1 Robustness of the System. The robustness of the system is an important factor affecting its use by BLV users. In practical use, illumination conditions, the color of the table and floor, and the number of target objects in the picture will affect object recognition, hand pose recognition and depth calculation. Therefore, when training the object detector, we collected pictures of different environments and augmented the pictures to a certain extent. According to the evaluation results, the object detector can recognize objects under most lighting conditions and with different table or floor colors. The training set of the hand pose recognizer contains the pictures collected in the field environment. The training set includes complex backgrounds and dynamic data augmentation [76]. Therefore, there are few errors in hand recognition. However, it is worth noting that object recognition, hand pose recognition and depth calculation still fail to work in the case of strong illumination. We believe that using more advanced cameras or image preprocessing can further alleviate this problem.

7.2.2 Expanding Categories of Object Recognition. At present, we have only completed the identification of 10 common objects in BLV users' daily lives. However, these classifications are not fine enough. Moreover, the objects that BLV users need to reach are not necessarily already registered. In order to improve practicability, it is necessary to retrain the personal object detector to recognize more categories. According to the analysis discussed in Section 6.1.5, we found that although we can augment the data of the new training set to obtain more training samples, sufficient original training samples (20 samples per object category) are the premise to ensure the recognition accuracy of the object detector. Moreover, our system is suitable for all day wear and can capture useful videos all day. Therefore, we can ease the user's registration burden through the automatic discovery of handheld objects for further analysis and processing [52, 71, 79].

7.2.3 Loss of the Target Object. We found that BLV users put their hands on the table or floor to reach the target object, which does not block the target object most of the time. However, when the hand approaches the target object (especially the tiny key), their hands sometimes block the target object, which will cause the object to be lost in some frames. A potential method to overcome this issue is to an embed Inertial Measurement Unit (IMU) in the headphone to predict the user's movement and then predict the target position [80]. However, although we use a large wide-angle lens module, the camera field is still not large enough. We found that in the case of things fallen on the floor, BLV users need to crouch down to reach the object, which makes the distance between the camera and the floor shorter, which leads to the narrowing of the field of vision, thus causing the loss of the object. One possible solution is to use a wider lens. In the future, we will consider using a flat angle lens with a larger field. It is worth mentioning that the increase in the field of view means more pixels should be processed, which further impacts the delay.

7.2.4 External Server. Similar to many current mobile assistive devices for BLV users [22, 23], LiSee senses environmental information on a wearable mobile terminal and uses external devices or servers to process information. At present, most computing intensive mobile systems adopt this external information processing method because of the limitations of battery capacity and computing power of the mobile terminal [41, 51]. However, we believe that the battery capacity and computing power of mobile terminals will be improved in the future. With the development of Wireless Power Transmission, the transmission distance and efficiency are constantly improving [58], which makes wireless charging possible to solve the battery capacity problem of LiSee in the future. With the advent of Application-Specific Integrated Circuit (ASIC) (such as TPU [7]), it is possible to integrate special processing units into wearable devices. ASIC can be customized for specific algorithms to realize the advantages of small volume, low power consumption, and high computing performance and efficiency. Nevertheless, once the customized ASIC is manufactured, it cannot be changed, so the initial cost of research and development is high and the development cycle is long. We envisage customizing special ASICs in the future to remove external servers and improve LiSee's mobility.

7.3 Interactivity

7.3.1 More Efficient Guidance. In the present study, we pay more attention to the common design process and universality of LiSee. Therefore, we provide personalized guidance options for participants based on their suggestions, so that they can quickly learn to use LiSee based on their own knowledge. In fact, personalized selection and design [28, 35] are the current development trend. However, there are fewer participants involved in co-design and more guidance options (although training may be required) have not been explored. In the future, our aim is to recruit more BLV users, design more guidance programs, and explore optimal guidance programs through more user experiments.

7.3.2 More Natural Interactivity. Our speech guide scheme is concise and compact and designed to provide guidance quickly, but some users (P3, P7, P12) find it rigid. We consider that deep learning and natural language understanding can provide more natural interactions, which may be useful in improving the user experience. With the research and progress in the field of natural language understanding, many algorithms show excellent performance [33, 75], which provides a reference for our future work.

7.4 Preliminary Study with Few BLV Users

As a preliminary study, our goal was to design and evaluate LiSee with BLV users. However, because the co-design process requires in-depth interaction and evaluation with the BLV users, the study sample size was small. Indeed, the small sample size limits the generalizability of the proposed system. However, we note that the difficulties and design requirements of the BLV users participating in our design were similar, and LiSee showed a considerable

advantage over not using LiSee across all participants. Therefore, we are confident that LiSee will maintain similar advantages across a wider range of evaluations. Nevertheless, we will still consider conducting evaluations with larger sample sizes in the future.

7.5 Beyond Reaching Target Objects

We propose a novel wearable neckband headphone, which is suitable for BLV users to wear for a long time and use frequently. In addition, it is unobtrusive and esthetic. Compared with the limited holding space of glasses [31, 62], the neckband headphone system we propose includes hardware such as a binocular camera, microphone, and speaker, which provides certain scalability. As is well known, BLV users encounter many difficulties in daily life, such as navigation, obstacle avoidance, and text recognition. Based on our system hardware, modifications at the software level can provide additional assistance. While we demonstrated the excellent performance of the neckband headphone in reaching target objects, there may be many new challenges to overcome in other tasks. For example, when BLV users walk to avoid obstacles, motion blur caused by the instability of the camera may become more obvious. We believe that integrating multiple features on an all-day wearable and ready-to-use system can make life easier for BLV users.

8 CONCLUSION

We proposed a lossless wearable system, the first neckband headphone wearable system that provides all-day assistance for BLV users to reach surrounding objects. We used a user-centered design method and investigated the difficulties and requirements of BLV users through interviews. According to the interviews, we further designed a novel neckband headphone form, taking the function of the system as an expansion of the existing headphone, which is suitable for BLV users to wear all day and use at any time in their daily life. In order to adapt to the new form, we also designed a set of seamless image processing algorithms, and used cloud and fog collaborative computing to meet the computational power requirements. In order to enable users to quickly use the system based on their personal expertise, we provided personalized guidance schemes for users to choose. Finally, the studies in the laboratory and participants' homes showed that LiSee can successfully guide users to reach target objects and meet their daily needs. As a promising system, the combination of acceptability, wearability, comfort, and esthetics has been raised to a new level. We believe that with the continuous technical iteration, the system can enter the daily lives of BLV users.

ACKNOWLEDGMENTS

We would like to thank all the participants for their support. We would also like to thank the anonymous reviewers for their valuable suggestions. The research was supported in part by the China NSFC 61872246. Lu Wang is the corresponding author.

REFERENCES

[1] 2007. VOC2007. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html.
[2] 2012. VOC2012. http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html.
[3] 2022. Apple ARKit. https://developer.apple.com/cn/augmented-reality/arkit/.
[4] 2022. BeMyEyes. https://www.bemyeyes.com/.
[5] 2022. Envision AI. https://apps.apple.com/us/app/envision-ai/id1268632314.
[6] 2022. Google Glass. https://www.google.com/glass/start/.
[7] 2022. Google TPU. https://cloud.google.com/tpu/docs/tpus.
[8] 2022. Help People who are Blind or Partially Sighted. https://www.orcam.com/en/.
[9] 2022. iFLYTEK. https://www.iflytek.com/.
[10] 2022. IT9100. https://www.itech.sh/en/product/power-meter/IT9100.html.
[11] 2022. KNFB. https://knfbreader.com/.

[12] 2022. LookTel Recognizer. https://http://www.looktel.com/recognizer.
[13] 2022. Seeing AI. https://www.microsoft.com/en-us/ai/seeing-a.
[14] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.
[15] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. 2020. " I am uncomfortable sharing what I can't see": Privacy Concerns of the Visually Impaired with Camera Based Assistive Applications. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 1929–1948.
[16] Swamy Ananthanarayan, Miranda Sheh, Alice Chien, Halley Profita, and Katie Siek. 2013. Pt Viz: towards a wearable device for visualizing knee rehabilitation exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1247–1250.
[17] World Medical Association et al. 2013. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *Jama* 310, 20 (2013), 2191–2194.
[18] Mauro Avila Soto and Markus Funk. 2018. Look, a guidance drone! assessing the social acceptability of companion drones for blind travelers in public spaces. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 417–419.
[19] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies* (2009).
[20] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-up robust features (SURF). *Computer vision and image understanding* 110, 3 (2008), 346–359.
[21] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.
[22] Roger Boldu, Alexandru Dancu, Denys JC Matthies, Thisum Buddhika, Shamane Siriwardhana, and Suranga Nanayakkara. 2018. Fingerreader2. 0: Designing and evaluating a wearable finger-worn camera to assist people with visual impairments while shopping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–19.
[23] Roger Boldu, Denys JC Matthies, Haimo Zhang, and Suranga Nanayakkara. 2020. AiSee: An Assistive Wearable Device to Support Visually Impaired Grocery Shoppers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–25.
[24] Stacy M Branham and Shaun K Kane. 2015. Collaborative accessibility: How blind and sighted companions co-create accessible home spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2373–2382.
[25] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[26] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996).
[27] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *The 25th annual international conference on mobile computing and networking*. 1–17.
[28] Y. Chang, Y. Zhao, M. Dong, Y. Wang, and L. Shang. 2021. MemX: An Attention-Aware Smart Eyewear System for Personalized Moment Auto-capture. *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–23.
[29] Wenqiang Chen, Lin Chen, Yandao Huang, Xinyu Zhang, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2019. Taprint: Secure text input for commodity smart wristbands. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.
[30] Xiaogang Chen, Jie Yang, Qiang Wu, and Jiajia Zhao. 2010. Motion blur detection based on lowest directional high-frequency energy. In *2010 IEEE International Conference on Image Processing*. IEEE, 2533–2536.
[31] Patrick Chwalek, David Ramsay, and Joseph A Paradiso. 2021. Captivates: A Smart Eyeglass Platform for Across-Context Physiological Measurement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–32.
[32] Artem Dementyev and Christian Holz. 2017. DualBlink: A Wearable Device to Continuously Detect, Track, and Actuate Blinking For Alleviating Dry Eyes and Computer Vision Syndrome. *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 1, 1 (2017), 1–19.
[33] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems* 32 (2019).
[34] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.
[35] Steven M Goodman, Ping Liu, Dhruv Jain, Emma J Mcdonnell, Jon E. Froehlich, and Leah Findlater. 2021. Toward User-Driven Sound Recognizer Personalization with People Who Are d/Deaf or Hard of Hearing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2021).
[36] Nadia Gosselin-Kessiby, John F Kalaska, and Julie Messier. 2009. Evidence for a proprioception-based rapid on-line error correction mechanism for hand orientation during reaching movements in blind subjects. *Journal of neuroscience* 29, 11 (2009), 3485–3496.
[37] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. 2013. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics* 43, 5 (2013), 1318–1334.

[38] Rachel Hewett, Graeme Douglas, and Sue Keil. 2015. Young people, visual impairment and preparing to live independently. *Visual Impairment Centre for Teaching and Research, University of Birmingham* (2015).
[39] Heiko Hirschmuller. 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 807–814.
[40] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7310–7311.
[41] Yongzhi Huang, Kaixin Chen, Lu Wang, Yinying Dong, Qianyi Huang, and Kaishun Wu. 2021. Lili: liquor quality monitoring based on light signals. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 256–268.
[42] Felix Huppert, Gerold Hoelzl, and Matthias Kranz. 2021. GuideCopter-A Precise Drone-Based Haptic Guidance Interface for Blind or Visually Impaired People. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[43] Watthanasak Jeamwatthanachai, Mike Wald, and Gary Wills. 2019. Indoor navigation by blind people: Behaviors and challenges in unfamiliar spaces and buildings. *British Journal of Visual Impairment* 37, 2 (2019), 140–153.
[44] Glenda Jessup, Anita C Bundy, Alex Broom, and Nicola Hancock. 2018. Fitting in or feeling excluded: The experiences of high school students with visual impairments. *Journal of Visual Impairment & Blindness* 112, 3 (2018), 261–273.
[45] Lingqiu Jin, He Zhang, Yantao Shen, and Cang Ye. 2020. Human-Robot Interaction for Assisted Object Grasping by a Wearable Robotic Object Manipulation Aid for the Blind. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–6.
[46] Shaun K Kane. 2019. Wearables. In *Web accessibility*. Springer, 701–714.
[47] Oliver Beren Kaul, Kersten Behrens, and Michael Rohs. 2021. Mobile Recognition and Tracking of Objects in the Environment through Augmented Reality and 3D Audio Cues for People with Visual Impairments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.
[48] Vinitha Khambadkar and Eelke Folmer. 2013. GIST: a gestural interface for remote nonvisual spatial perception. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 301–310.
[49] Wonjung Kim, Seungchul Lee, Seonghoon Kim, Sungbin Jo, Chungkuk Yoo, Inseok Hwang, Seungwoo Kang, and Junehwa Song. 2020. Dyadic Mirror: Everyday Second-person Live-view for Empathetic Reflection upon Parent-child Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.
[50] Boon Tatt Koik and Haidi Ibrahim. 2013. A literature survey on blur detection algorithms for digital imaging. In *2013 1st International Conference on Artificial Intelligence, Modelling and Simulation*. IEEE, 272–277.
[51] Karthik Kumar, Jibang Liu, Yung-Hsiang Lu, and Bharat Bhargava. 2013. A survey of computation offloading for mobile systems. *Mobile networks and Applications* 18, 1 (2013), 129–140.
[52] Franklin Mingzhe Li, Di Laura Chen, Mingming Fan, and Khai N Truong. 2019. FMT: A wearable camera-based object tracking memory aid for older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.
[53] Jun Li, Reinhard Klein, and Angela Yao. 2017. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*. 3372–3380.
[54] Tianxing Li and Xia Zhou. 2018. Battery-free eye tracker on glasses. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 67–82.
[55] Xiaoping Liu, He Zhang, Lingqiu Jin, and Cang Ye. 2018. A wearable robotic object manipulation aid for the visually impaired. In *2018 IEEE 1st International Conference on Micro/Nano Sensors for AI, Healthcare, and Robotics (NSENS)*. IEEE, 5–9.
[56] Robyn Longhurst. 2003. Semi-structured interviews and focus groups. *Key methods in geography* 3, 2 (2003), 143–156.
[57] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.
[58] Xiao Lu, Ping Wang, Dusit Niyato, Dong In Kim, and Zhu Han. 2015. Wireless charging technologies: Fundamentals, standards, and network applications. *IEEE communications surveys & tutorials* 18, 2 (2015), 1413–1452.
[59] Hac Maruri, P. Lopez-Meyer, J. Huang, W. M. Beltman, and L. Hong. 2018. V-Speech: Noise-Robust Speech Capturing Glasses Using Vibration Sensors. *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.
[60] Karsten Mühlmann, Dennis Maier, Jürgen Hesser, and Reinhard Männer. 2002. Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision* 47, 1 (2002), 79–88.
[61] Rahul Nair, Kai Ruhl, Frank Lenzen, Stephan Meister, Henrik Schäfer, Christoph S Garbe, Martin Eisemann, Marcus Magnor, and Daniel Kondermann. 2013. A survey on time-of-flight stereo fusion. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 105–127.
[62] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-day Captioning. In *UIST '20: The 33rd Annual ACM Symposium on User Interface Software and Technology*.
[63] World Health Organization et al. 2004. *ICD-10: international statistical classification of diseases and related health problems: tenth revision*. World Health Organization.

[64] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.
[65] H. Profita, R. Albaghli, L. Findlater, P. Jaeger, and S. K. Kane. 2016. The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use. *ACM* (2016).
[66] Krishna Ribeiro-Gomes, David Hernandez-Lopez, Rocío Ballesteros, and Miguel A Moreno. 2016. Approximate georeferencing and automatic blurred image detection to reduce the costs of UAV use in environmental and agricultural applications. *Biosystems Engineering* 151 (2016), 308–327.
[67] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*. Ieee, 2564–2571.
[68] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.
[69] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.
[70] Shantanu A Satpute, Janet R Canady, Roberta L Klatzky, and George D Stetten. 2019. FingerSight: A Vibrotactile Wearable Ring for Assistance With Locating and Reaching Objects in Peripersonal Space. *IEEE transactions on haptics* 13, 2 (2019), 325–333.
[71] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. 2020. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9869–9878.
[72] Meng-Li Shih, Yi-Chun Chen, Chia-Yu Tung, Cheng Sun, Ching-Ju Cheng, Liwei Chan, Srenivas Varadarajan, and Min Sun. 2018. Dlwv2: A deep learning-based wearable vision-system with vibrotactile-feedback for visually impaired people to reach objects. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1–9.
[73] Kevin V Stefanik, Jason C Gassaway, Kevin Kochersberger, and A Lynn Abbott. 2011. UAV-based stereo vision for rapid aerial terrain mapping. *GIScience & Remote Sensing* 48, 1 (2011), 24–49.
[74] Nelson Daniel Troncoso Aldas, Sooyeon Lee, Chonghan Lee, Mary Beth Rosson, John M Carroll, and Vijaykrishnan Narayanan. 2020. AIGuide: An Augmented Reality Hand Guidance Application for People with Visual Impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13.
[75] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and Samuel R Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. (2018).
[76] Yangang Wang, Baowen Zhang, and Cong Peng. 2019. Srhandnet: Real-time 2d hand pose estimation with simultaneous region localization. *IEEE transactions on image processing* 29 (2019), 2977–2986.
[77] Fredrik Winberg and John Bowers. 2004. Assembling the senses: towards the design of cooperative interfaces for visually impaired users. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. 332–341.
[78] Wentao Xie, Qian Zhang, and Jin Zhang. 2021. Acoustic-based Upper Facial Action Recognition for Smart Eyewear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–28.
[79] Takuma Yagi, Takumi Nishiyasu, Kunimasa Kawasaki, Moe Matsuki, and Yoichi Sato. 2021. GO-finder: a registration-free wearable system for assisting users in finding lost objects via hand-held object discovery. In *26th International Conference on Intelligent User Interfaces*. 139–149.
[80] Zhijian Yang, Yulin Wei, Sheng Shen, and Romit Roy Choudhury. 2020. Ear-AR: indoor acoustic augmented reality on earphones. In *MobiCom '20: The 26th Annual International Conference on Mobile Computing and Networking*.
[81] Chien Wen Yuan, Benjamin V Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M Carroll. 2019. Constructing a holistic view of shopping with people with visual impairment: a participatory design approach. *Universal Access in the Information Society* 18, 1 (2019), 127–140.
[82] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.
[83] Shibo Zhang, Yuqi Zhao, Dzung Tri Nguyen, Runsheng Xu, Sougata Sen, Josiah Hester, and Nabil Alshurafa. 2020. Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–26.
[84] Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. 2016. CueSee: exploring visual cues for people with low vision to facilitate a visual search task. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 73–84.
[85] Peter A Zientara, Sooyeon Lee, Gus H Smith, Rorry Brenner, Laurent Itti, Mary B Rosson, John M Carroll, Kevin M Irick, and Vijaykrishnan Narayanan. 2017. Third eye: A shopping assistant for the visually impaired. *Computer* 50, 2 (2017), 16–24.