StereoPilot: A Wearable Target Location System for Blind and Visually Impaired Using Spatial Audio Rendering

Xuhui Hu, *Graduate Student Member, IEEE*, Aiguo Song, *Senior Member, IEEE*, Zhikai Wei, *Student Member, IEEE*, and Hong Zeng, *Member, IEEE*

***Abstract*— Vision loss severely impacts object recognition and spatial cognition for limited vision individuals. It is a challenge to compensate for this using other sensory modalities, such as touch or hearing. This paper introduces StereoPilot, a wearable target location system to facilitate the spatial cognition of BVI. Through wearing a head-mounted RGB-D camera, the 3D spatial information of the environment is measured and processed into navigation cues. Leveraging spatial audio rendering (SAR) technology, it allows the navigation cues to be transmitted in a type of 3D sound from which the sound orientation can be distinguished by the sound localization instincts in humans. Three haptic and auditory display strategies were compared with SAR through experiments with three BVI and four sighted subjects. Compared with mainstream speech instructional feedback, the experimental results of the Fitts' law test showed that SAR increases the information transfer rate (ITR) by a factor of three for spatial navigation, while the positioning error is reduced by 40%. Furthermore, SAR has a lower learning effect than other sonification approaches such as vOICe. In desktop manipulation experiments, StereoPilot was able to obtain precise localization of desktop objects while reducing the completion time of target grasping tasks in half as compared to the voice instruction method. In summary, StereoPilot provides an innovative wearable target location solution that swiftly and intuitively transmits environmental information to BVI individuals in the real world.**

***Index Terms*— Blind assistance, environment perception, sensory feedback, spatial audio.**

I. Introduction

VISION is commonly considered as the primary sensory modality for spatial cognition and object recognition [1], [2]. Globally, an estimated 43.3 million people were considered blind in 2020 [3]. People with blindness or visual impairment (BVI) not only compromise their autonomy in many daily living activities, such as selecting color matching clothes or identifying a desired product in the supermarket, but also leads to lower work-force participation rates and higher risks of falls. The development of computer vision (CV) technology opens the potential for BVI individuals to obtain environmental information, but there is still no neural-like perceptual feedback approach that can quickly and intuitively convey high-bandwidth environmental information to users. The motivation of this paper is to improve the information transfer rate (ITR) from the environment to BVI, so as to facilitate their spatial cognition of the environment.

There are a wide variety of assistance technologies that can help BVI with their daily activities. "Argus II Retinal Prosthesis System" is one of the neural implants technologies developed by Second Sight Inc. [4], it uses an external camera that linked to a tiny implant array of 60 electrodes attached to the retina. Argus II can restore rudimentary vision of BVI, however, the image information delivered by the implant electrodes is quite limited, the high risks and high costs of invasive treatments have also hindered its widespread application. Therefore, most blind assistance strategies utilize the remaining sensory neural pathways of the target population to transmit information, e.g. BVI's haptic and auditory abilities. Currently, the primary presentation modality is auditory display, that is using sound to communicate the information of the machine to users [5], [6]. For instance, Helal *et al.* [7] developed a wireless pedestrian navigation system for BVI, where users can interact with the system through voice instruction (VI) and the environmental information is provided through detailed speech cues. Recently, Microsoft Inc. developed "Seeing AI", a CV based blind-aid software for smartphones. It can identify people and objects through the device camera, and then the app can audibly describe those objects by synthetic speech. It supports multiple practical tasks such as identifying short text, documents, products, person, scene, currency and so on. Furthermore, Meijer invented a novel image-to-sound rendering technology termed "vOICe" [8], through which the time sequence of the sound wave, the pitch, and the loudness are mapped to the row, the column and the brightness of the image respectively, offering an experience of live camera views for BVI.

Haptic display uses mechanical actuators or electrical stimulators to provide a sense of touch, which allows BVIs to

Manuscript received February 15, 2022; revised May 8, 2022 and June 7, 2022; accepted June 9, 2022. Date of publication June 13, 2022; date of current version June 21, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 92148205 and Grant 62173089, in part by the Basic Research Program of Jiangsu Province under Grant BK201900240, and in part by the Basic Research Funds for Universities under Grant 2242022K30056. *(Corresponding author: Aiguo Song.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of the University and performed in line with the declaration of Helsinki.

The authors are with the State Key Laboratory of Bioelectronics and the Jiangsu Key Laboratory of Remote Measurement and Control, School of Instrument Science and Engineering, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: a.g.song@seu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3182661

feel textures and shapes of virtual objects. There are variety of haptic display approaches for different applications: Brewster *et al.* [9], [10] combined a haptic device with non-speech sound cues to provide BVIs with a richer and more flexible form of access to graphs and tables. Ni *et al.* [11], [12] developed a walking assistance robot based on vibrotactile perception; Ulrich *et al.* [13] developed a hand-held mobile cane for walking assistance; Bonani *et al.* [14] developed a human-machine collaborative robot system based on the Baxter robot platform, which provides arm guidance to manipulate close distance objects; Chen *et al.* [15] built an vibrotactile-based image contour display system for BVI to access images conveniently through the touch screen. In addition, Sampaio *et al.* [16] developed an electro-tactile tongue display unit (TDU) with a 12×12 stainless steel electrodes array, it's commercial version is known as BrainPort (Wicab Inc.) [17]. TDU stimulates the perception area of tongue, so it cannot eat or talk while using the tongue display.

Although the above technologies have played an important role in environmental information recognition and intelligent human-computer interaction for BVI individuals, these methods have certain limitations in environmental navigation tasks: On the one hand, users need hands to interact with the environment, so they cannot use hand-held force feedback devices to transmit information. On the other hand, mainstream auditory display technologies, such as speech, sonification, and earcons, are not spatialized [18], thus it is time-consuming and non-intuitive to transmit spatial information. 3D audio technology transforms synthetic sound waves into simulated natural sound waves emanating from a point in 3D space. It allows trickery of the brain using the ears and auditory nerves, pretending to place a sound source in 3D space upon hearing the sounds produced by a pair of earphones. Frauenberger *et al.* [19] proposed a 3D sound-based human-computer interaction system to increase the information flow between a computer and the user. Walker *et al.* [20] developed "SWAN" for safe pedestrian navigation based on 3D sonification. Recently, Microsoft Inc. has also developed "Soundscape", it uses 3D audio cues to enrich ambient awareness with the help of high-precision map information.

The problem space of the above studies is mainly focused on outdoor navigation, which uses Global Positioning System (GPS) to obtain environmental information, however, indoor navigation has been less well supported by such information technology. May *et al.* [21] investigated the effectiveness of area cues and proximity feedback in facilitating object targeting based on virtual environment, but to our best knowledge there is rarely research on BVI's hand prehension task in real environment. The problem space of this paper focuses on object recognition and localization in desktop scenes. We introduce a notion that leverages spatial audio rendering (SAR) to transmit environmental information, which contributes to a faster ITR and improves users' spatial cognition with limited visual input. We used a wearable RGB-D camera to obtain indoor environmental information, then the navigation cues were encoded into stereo sound cues, from which users can easily perceive orientation information through the instinct of sound localization.

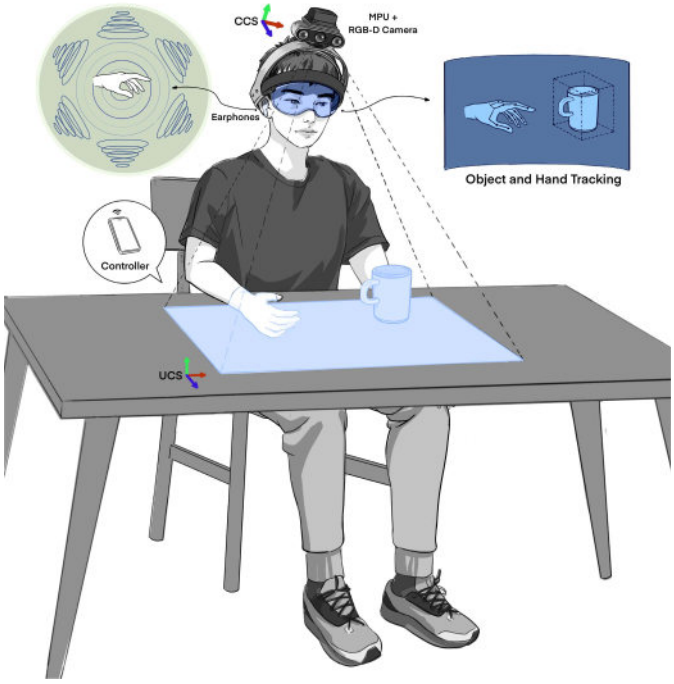

Fig. 1. The design concept of StereoPilot.

Our previous study used a fixed-position camera to identify objects and conducted preliminary experiments on sighted subjects [22]. This study involved in-depth research on the wearable design of an assistance device and extensive comparative experiments on target populations. The contributions of this paper include the following: 1) The feasibility and positioning accuracy of the wearable visual perception module were tested. 2) The ITR for the SAR on BVI was evaluated and compared with three other baseline feedback strategies based on auditory display and haptic display methods. 3) As an evaluation task, the developed assistance device was tested under the functional domain of daily life (practical life skills), in which the target objects must be detected, identified, and manipulated.

II. Methods

This section presents the system framework and typical workflow for StereoPilot. We introduce the environment perception method based on a wearable camera and the method of environmental information feedback based on SAR.

A. System Framework

The concept of the StereoPilot is depicted in Fig. 1. We present a wearable visual perception module that identifies and locates objects in the environment. Stereo earphones provide both VI-based object recognition feedback and SAR-based location assistance. Object recognition was performed using a head-mounted RGB-D camera, which allowed the tracking algorithm to obtain the 3D position of objects under the camera's coordinate system. In order to obtain the spatial information consistent with a user's spatial cognition, we estimate the user's head posture to transform the coordinates from the camera coordinate system (CCS) to the user coordinate system (UCS). To provide intuitive

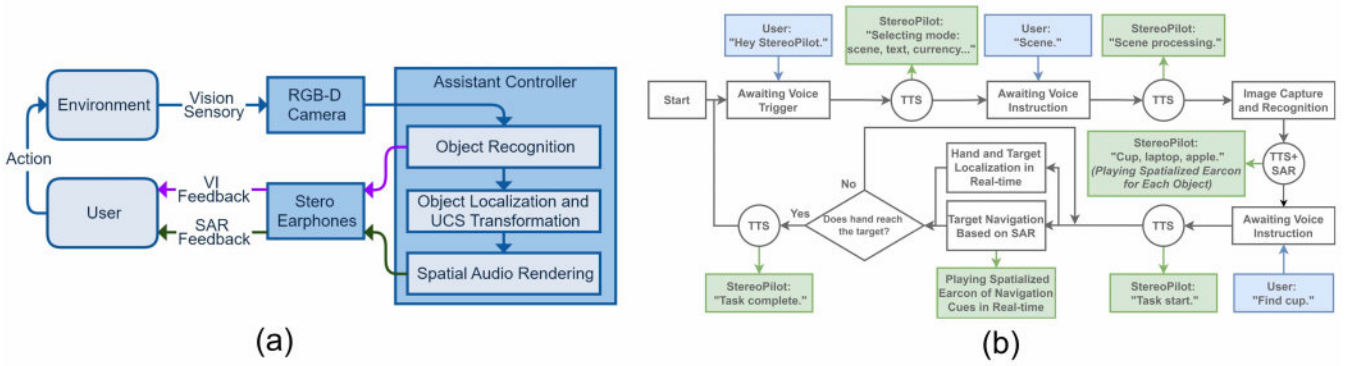

Fig. 2. (a) The system framework and (b) the typical workflow of StereoPilot. (TTS: text to speech; SAR: spatial audio rendering).

target location for BVI, we propose a spatialized auditory feedback method using spatial audio rendering. We designed a virtual reality scene that reflects the spatial information of the environment in real-time, then positioned a virtual sound source with spatial audio effects on a target object in the virtual world. Therefore, by perceiving the spatial audio generated in a virtual environment, the user would produce the illusion that the target object is vocalizing in the real environment. Then, the user only needs to follow the sound cues to achieve localization assistance. To meet the daily environmental perception requirements of BVI, the following technical considerations are included:

1) The camera should have a wide field of view (FoV) and a depth perception ability to handle both the long-distance environmental perception used for mobile assistance [23] and the short-distance environmental perception used for object-localization assistance. We used the RealSense D435 Depth Camera (Intel Inc.) with FoVs for the integrated RGB sensor of $69° \times 42°(H \times V)$, and for the depth sensor of $87° \times 58°(H \times V)$. The depth sensory range reaches $10m$.

2) The blind-aid device should have a wearable design. Existing visual assistance devices often integrate cameras into glasses or helmets. Compared with other fixed methods, such as a chest mount harness, the advantages of the head-mounted design are as follows. First, it can obtain a better environment perception perspective. Second, the visual range can be easily expanded by rotating the head without laboriously moving the torso. Third, it is more convenient for BVI to wear a helmet than a chest-mounted harness.

The challenge brought by using head-mounted cameras is that the CCS changes with head movements, which makes it difficult to calibrate spatial information. Thus, we designed a head-mounted visual perception helmet equipped with a D435 camera and a 9-axis motion processing unit (MPU9250, TDK Inc.), as shown in Fig. 1. The MPU9250 can estimate the posture of the camera in real-time, which can transform environmental information from CCS to UCS. Additionally, the user needs to wear a pair of stereo earphones (AirPods Pro, Apple Inc.) to obtain audio feedback information. We used a smart mobile device as the StereoPilot controller (e.g. smartphone, tablet, etc.) to improve the portability of the assistance robot.

The system framework of the StereoPilot is shown in Fig. 2(a). Firstly, the RGB-D camera transmits the environmental video stream to the assistant controller for object and hand recognition. Then, the recognition result is sent back to the user via voice instruction (VI) feedback. We use object localization and coordinate system transformation algorithms to obtain the spatial information under UCS, which is then fed into the virtual environment to generate spatial audio, finally received by the user through the stereo earphones. Therefore, the StereoPilot forms a closed control loop that enables users to interact with the environment in real-time. The StereoPilot was designed to handle close-distance environmental perception, but the framework is also suitable for long distances (e.g., mobile navigation).

A typical workflow for StereoPilot is shown in Fig. 2(b), here we focus on environment recognition and target location tasks. First, the user says "Hey StereoPilot" as a voice trigger to invoke the system, then the system guides the user to select corresponding modes through synthetic speech, including but not limited to those shown in the dialog block (e.g. scene, text, and currency recognition). Second, the user says "Scene" according to the conversation, and the system will prompt "Scene processing" to confirm the task status to the user. Then the system captures a photo through the RGB-D camera for general object recognition and says the recognition results. Each time an object is spoken, a spatialized earcon will be emitted simultaneously to indicate the location of the object relative to the camera. Third, the user can further specify the task based on the recognition results, e.g. "Find cup", then the system will prompt "Task start" and activate the target location via spatial audio feedback. The user will hear the spatialized navigation cues of a target object relative to the hand, the specific feedback form of SAR is described in Section II-C. When the system identifies that the coordinates of the hand coincide with the target coordinates (the user has grabbed the object), the system will prompt "Task complete" and then terminate current mode, the user can use other modes by saying the voice trigger command again. In addition, the user can say "Cancel" to quit the running mode. The StereoPilot interacts primarily with users through voice, so the hands are free to perceive the environment instead of operating the assistance device.

B. Spatial Perception Based on Computer Vision

Thanks to advances in artificial intelligence and innovations in deep learning and neural networks, computer vision has

taken great leaps in recent years and has surpassed humans in some tasks related to object detection and labeling [24], [25]. However, the vast number of calculations and power consumption are the main bottlenecks that limit its application in wearable devices. Here, we used the MediaPipe (Google Inc.) framework to identify and track objects and hands. MediaPipe is based on the most cutting-edge machine learning solutions, and the advantages of end-to-end acceleration are easily deployed on mobile devices, such as Android or iOS smartphones [26]. We used NucBox S (GMK Inc.) as the assistant controller as it is a pocket-sized x86 hardware development platform ($6cm \times 6cm \times 4cm$, $122g$). The device has an Intel J4125 (maximum frequency 2.7 GHz, 4-Core and 4-Thread) CPU, with 256 GB SSD and 8 GB LPDDR4 RAM, which has an equivalent performance to mainstream smart mobile devices. We used NucBox to develop the RealSense camera and MediaPipe on the Windows platform, where the D435 camera and MPU module are wire connected to the controller.

We first obtained the pixel coordinates of the desktop objects and the hand in the RGB image based on MediaPipe's recognition framework. The RGB image was then aligned with the point cloud map of the depth sensor to obtain the 3D positions of objects and hands in the CCS. To obtain the location information consistent with typical spatial perception, we calculated the quaternions of the MPU ($Q = q_0 + q_1 i + q_2 j + q_3 k$) in real-time to estimate the rotation transformation matrix of the camera in Eq. 1, as shown at the bottom of the page, which converts the 3D point cloud data of the CCS into the x_{ucs} coordinates of the UCS, as shown in Eq. 2, as shown at the bottom of the page.

The definitions of the UCS and CCS are shown in Fig. 1. The UCS is aligned with the upright posture of a human, where the x-axis corresponds to the left-right axis of the human body, the y-axis corresponds to the dorsoventral axis, and the z-axis corresponds to the craniocaudal axis, which is consistent with the user's spatial cognition.

C. Target Location Based on Spatial Audio Rendering

The human auditory system has inborn spatial perception abilities, this is due to the unique structure of the ear. The human pinna (external ear) modulates sounds before reaching the inner ear, which changes the sound's incoming path, arrival time, energy loss, etc. Thus, the auditory nerve can distinguish the orientation of a sound [27]–[29]. In addition, the sound reflection in the environment helps perceive the distance [27]. In summary, there are three design aspects that make individuals distinguish a sense of orientation from rendered spatial audio.

1) Sound source design. We used a sound cue containing both high-frequency (12kHz to 15 kHz) and low-frequency (0.2Hz to 3 kHz) components. The maximum sound level is around 50 dB and is primarily concentrated in the low-frequency band, which is most sensitive to the human ear.

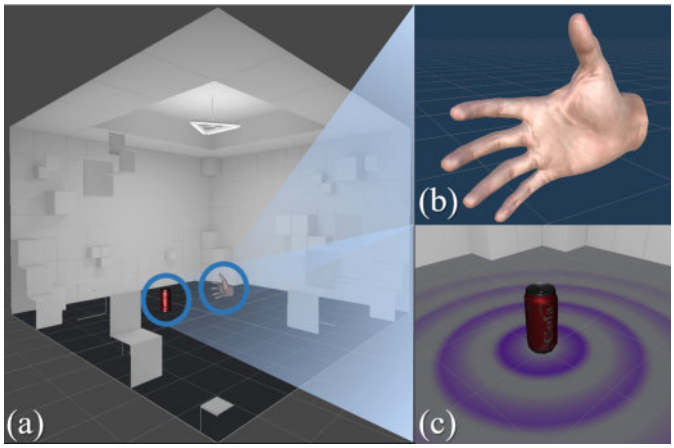

Fig. 3. The virtual environment for spatial audio rendering. (a) Audio room, (b) virtual hand, and (c) target object with virtual sound source.

2) Sound propagation model design. This can be solved using a head-related transfer function (HRTF) model, as shown in Eq. 3 and Eq. 4. We use this model to encode spatial information into sound. The specific parameters of the HRTF refer to the open-source database "CIPIC" [30], which offers a general HRTF model for normal-hearing subjects.

$$H_L(d,\theta,\varphi,\omega,\alpha) = X_L(d,\theta,\varphi,\omega,\alpha)/X_0(d,\omega) \quad (3)$$

$$H_R(d,\theta,\varphi,\omega,\alpha) = X_R(d,\theta,\varphi,\omega,\alpha)/X_0(d,\omega) \quad (4)$$

3) Sound reflection model design. We used "Resonance Audio" (Google Inc.) as the audio rendering engine and designed a virtual scene to render spatial audio based on the Unity platform (Version. 2019.4.15f1c1), as shown in Fig. 3(a). The target object and hand were placed in the sound room, their relative positions in the real environment were mapped into this virtual scene.We designed an enclosed cubic space in which the surface material, the sound source reflectance, and the reverberation properties of the sound room were fine-tuned to enhance sound reflections. The directivity pattern of the sound source was designed into a circular shape, which is represented as the purple circular effect in Fig. 3(c).

The general principle of SAR is to create a sound that seem to come from the same place as the target to which it refers. For some of the outdoor navigation systems based on 3D sound technology (e.g. soundscape and SWAN), the user directly acts as the moving object to perform the navigation task, so the ear is naturally used as the sound listener, as they can sense the position of the 3D sound by rotating the head or walking around the target. While in this study, the hand model located in the virtual scene was set as the sound listener, and the target located in the virtual scene was set

$$C_b^n = \begin{bmatrix} q_0^2+q_1^2-q_2^2-q_3^2 & 2(q_1q_2-q_0q_3) & 2(q_1q_3+q_0q_2) \\ 2(q_1q_2+q_0q_3) & q_0^2-q_1^2+q_2^2-q_3^2 & 2(q_2q_3-q_0q_1) \\ 2(q_1q_3-q_0q_2) & 2(q_2q_3+q_0q_1) & q_0^2-q_1^2-q_2^2+q_3^2 \end{bmatrix} \quad (1)$$

$$x_{ucs} = C_b^n \cdot x_{ccs} \quad (2)$$

TABLE I
SUBJECTS INFORMATION

Subject	Age (yrs.)	Gender	Vision
S1	21	Male	Sighted
S2	23	Female	Sighted
S3	28	Male	Sighted
S4	30	Male	Sighted
B1	21	Male	Congenital blindness
B2	23	Male	Congenital blindness
B3	24	Male	Congenital blindness

as the sound source. Therefore, the target position can be perceived more quickly and accurately by moving user's hand. Otherwise, when the sound listener is still set to the position of user's ear, the position between the ear and the object will not change much, thus the user cannot get effective spatial information from the target object. Our pilot experimental results show that, although this setup differs from the way humans perceive real sound sources, subjects only need a short period of training to get used to this navigation method.

III. EXPERIMENTS

This section compares SAR with three mainstream information feedback methods. We conducted ITR evaluation experiments to quantify the information feedback performance of different approaches, and designed desktop manipulation experiments to verify the feasibility of StereoPilot for accurate prehension tasks in real environment.

A. Subjects

Four sighted individuals (25.5 ± 4.2 years old) and three BVI individuals with congenital blindness (22.6 ± 1.5 years old) were informed and participated in the experiments. All subjects have normal hearing. The subject's information is shown in Table. I. Informed consent forms were obtained from all participants, and all experiments were approved by the ethical committee of the university and conformed to the declaration of Helsinki. During the experiments, the sighted subjects were blindfolded and could neither observe in advance nor perceive the environment with their hands.

B. Comparison of Spatial Information Feedback Methods

In order to make an unbiased comparison with the mainstream information feedback methods, we discussed the performance of three representative auditory and haptic display methods for the stated problem space. They are voice instruction feedback (termed VI group), vibrotactile feedback (termed VB group), and non-speech sonification feedback (termed NS group). Details of each feedback method are presented as follows:

1) VI is based on an intuitive notion that instruct the user with a small amount of voice cues. To avoid confusion caused by too many verbal directions, we used four commands: "forward", "backward", "left", and "right". The VI guides users through two steps: The subject must align the horizontal position before moving on to the vertical position. To depict the distance to the destination, we employed three distinct earcon frequencies (0.5 Hz, 1 Hz, and 2 Hz). The earcon frequency increased with increasing distance. To avoid auditory fatigue induced by high-frequency voice feedback, the direction speech instruction was played every six seconds. Details regarding the schematic of VI were presented in our earlier publication [22].

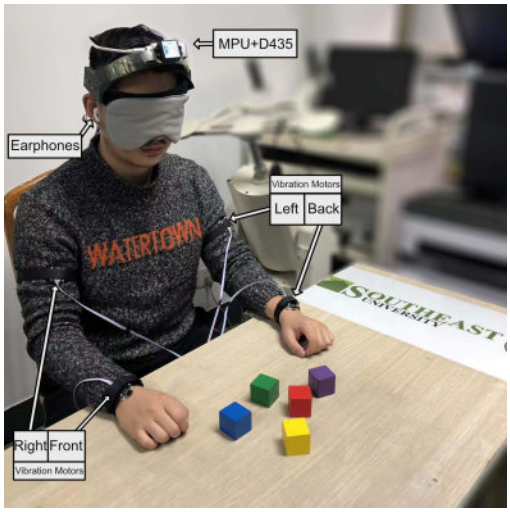

Fig. 4. The experimental setup of the desktop manipulation experiment.

2) VB encodes location information into vibration, and subsequently stimulates the tactile sensory nerves of human skin. Katzschmann *et al.* [31] placed five vibration motors in a linear and horizontal manner along an elastic strap worn across the upper abdomen. Kessler *et al.* [32] placed two vibration motors on the thenars of each hand and designed a two-tactor vibrotactile encoding strategy to get a continuous sense of direction, both methods achieve a forward sensing range of $\pm 90°$. To achieve an intuitive 360° sensing range with a minimum number of vibration motors, we designed a vibrotactile based spatial information feedback scheme using four vibration motors (as shown in Fig. 4) after evaluating the sensitivity of the user to the vibration: two of them were set on the biceps of both arms, representing the left-right direction, and the other two were set on the wrists of both, representing the anterior-posterior direction. The distance information is represented by the vibration intensity, which is modulated by the pulse width modulation (PWM) to the vibration motors, a stronger vibration intensity indicates being closer to the target. Details regarding vibrotactile coding have been presented in our pervious publication [33].

3) Numerous non-speech sonification (NS) approaches have been reported for blind-aid applications. Compared to other auditory display method, e.g. voice instruction feedback (VI), NS has a wider sound bandwidth to support a larger information capacity. Meanwhile, processing speech requires significant mental resources, as it is difficult to carry on a conversation while receiving speech cues. Mansur *et al.* [34] developed "Soundgraphs" to present line graphs in sound, where time line of the sound is mapped to the x-axis and pitch to the y-axis. The shape of the graph can then be heard as a rising or falling note playing over time. Here we

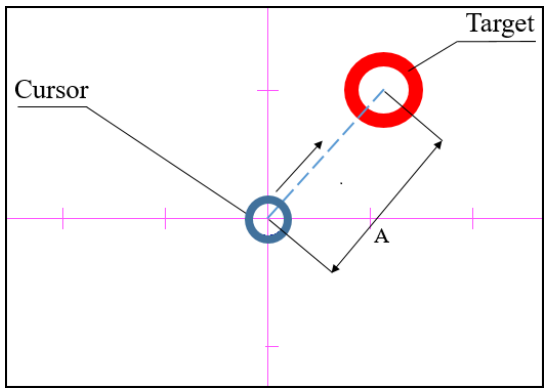

Fig. 5. The user interface of Fitts' law test.

used "vOICe" as the non-speech sonification benchmark, it is maintained by Meijer [8] and is freely accessible for research purposes. Similar to "Soundgraphs", video streams processed by "vOICe" are sounded in a left to right scan order, by default at a rate of one image snapshot per second. Therefore, hearing some sound from the left or right ear means a corresponding visual pattern on your left or right side, respectively. During every scan, pitch means elevation: the higher the pitch, the higher the position of the visual pattern. Loudness means brightness: the louder the brighter the image is. Consequently, silence means black, and a loud sound means white, and anything in between is grey.

C. Information Transfer Rate Evaluation Experiment

To quantify the ITR of each feedback method, we designed an evaluation experiment based on Fitts' law test. Fitts' law is primarily used in human–computer interactions and ergonomics evaluations, it expresses human movement as transmitting information through human channels [35], [36]. We designed a computer program for Fitts' law test, where the interface is shown in Fig. 5. There is a blue circle (cursor) and a fixed red circle (target) in a two-dimensional space, the experiment involves the user to move the cursor to the target point by manipulating the mouse based on the feedback information. The radius of the red target was 1.5 times that of the blue cursor. The task is considered as success only if the user presses the space bar while the blue cursor is inside the red target. In each trial, the target is randomly repositioned and the cursor always starts from the origin of the coordinate axis. In Fitts' law test, the ITR is defined as the throughput (TP), which is calculated based on the index of difficulty (ID) and the corresponding movement time (MT):

$$TP = \frac{1}{N}\sum_{i=1}^{N}\frac{ID_i}{MT_i} \tag{5}$$

$$ID = \ln\left(\frac{A}{\sqrt{2\pi e}\cdot\sigma} + 1\right) \tag{6}$$

where e is the Euler's number, A is the movement amplitude as represented by the initial distance between the cursor and the target. As the target contains a movement tolerance radius, σ represents the standard deviation of the positioning error between the final cursor and the target in all successful trials.

In SAR feedback, the sound source was set into the red circle, and the listener was set into the blue circle. While in NS feedback, We made some changes to the interface shown in Fig. 5, such as replacing the interface background with all black and keeping the target and cursor color and size the same. This was done to prevent vOICe from generating sound when scanning the blank content in the interface, thus relieving the user of unnecessary noise. Then the processed video stream of the program interface was directly transmitted into vOICe system, the snapshot update rate of vOICe is 1 fps. All the subjects performed SAR, VI, VB, and NS experiments in turn, and each group of experiments was repeated 30 times.

D. Desktop Manipulation Experiment

As described in Section III-B, the purpose of the Fitts' law test is to quantify the information transfer efficiency of different feedback methods in a normalized human-computer interface. In order to verify the feasibility of indoor target location tasks in real environment, this experiment focuses on the technology fusion of wearable visual perception and spatial information feedback. It should be noted that the position of the target object in the real environment and its physical properties have a significant impact on the user's grasping success rate. For instance, the closer the target object is to a group of interfering objects with similar tactile sensations, the easier it is to affect the user's judgement. However, when the physical properties of the target and the adjacent interfering object are significantly different (such as grasping one of the adjacent ceramic or paper cups), the user can easily distinguish them by hand, which increases the grasping success rate. Indeed, hands have been able to help BVI to complete many indoor localization tasks, but for some objects that only have visual differences, such as selecting specific colors of clothes or specific patterns of prints, BVI need effective technical support of computer vision and spatial information feedback.

To prevent biased experimental results caused by the above variables, the setup of the desktop manipulation experiment is shown in Fig. 4. The subject sits in front of the table, where there are five blocks consisting of the same material, mass, shape, but differed in color. The positions of all blocks were randomly scrambled by the experimenter, but the distance between each block is less than 5 cm. This is because too sparse distance makes the target object much easier to be located, which increases the grasping success rate.

In this experiment, subjects need to wear a visual perception helmet, and try to find the block with the specified color according to the guidance of the location information. First, the visual perception helmet uses the computer vision technology to identify the spatial information and the color of the blocks in the real environment, and the computer randomly assigns one of the colors as the color of the target object. Second, the system provides spatial location information of the target object to the subjects in real-time. Before each trial, the experimenters will shuffle the position of the blocks. Each subject need to complete 30 trials for each spatial information feedback method.

TABLE II
EVALUATION METRICS ON BVI AND SIGHTED SUBJECTS

Groups	BVI				SIGHTED			
	SAR	VI	VB	NS	SAR	VI	VB	NS
Error (a.u.)	0.047±0.031	0.065±0.029	0.077±0.029	0.083±0.030	0.060±0.033	0.074±0.033	0.086±0.034	0.082±0.027
Completion Time (s)	9.97±5.71	22.46±9.12	14.01±7.42	33.04±19.57	9.85±5.96	15.61±4.62	12.61±6.06	35.97±19.02
ITR (bits/s)	0.764	0.243	0.256	-0.596	0.471	0.329	0.367	-0.224
Pearson's r(a.u.)	0.232	0.457	0.400	-0.172	0.191	0.627	0.372	-0.0061
RMSE (a.u.)	5.970	6.828	6.484	19.240	4.995	3.238	4.855	19.470
Success Rate (%)	96.61±1.43	95.90±1.06	96.24±1.51	84.56±3.54	97.11±1.82	95.59±1.93	95.60±1.07	87.61±4.60

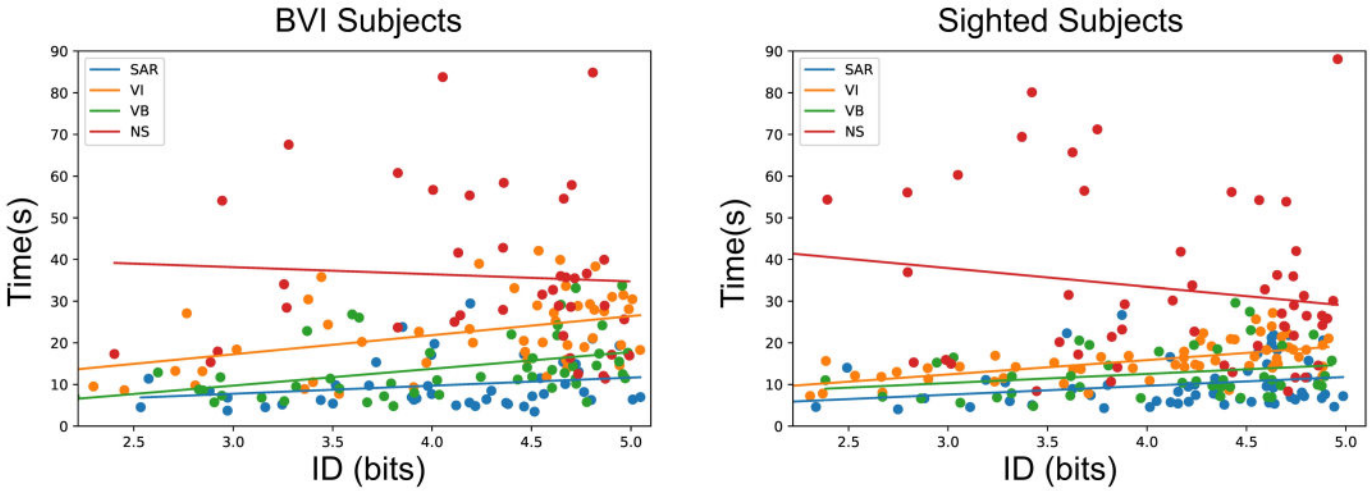

Fig. 6. Scatter plot based on MT and ID and the linear regression curve. For simplicity, only a portion of the sample points are shown.

IV. RESULTS

A. Fitts' Law Test

This subsection evaluates the performance of four feedback strategies in Fitts' law test, multiple metrics were used including positioning errors, completion time, ITR, Pearson correlation coefficient (Pearson's r) between the ID and MT, the root mean square error (RMSE) of the linear regression curve, and the success rate. The evaluation results on BVI and sighted subjects are shown in Table. II.

The scatter plot based on the MT and ID is shown in Fig. 6. The slope of the linear regression curve represents the ITR and shows that SAR was nearly three times that of the VI and VB in the BVI group. However, the NS group has a negative ITR, because subjects appeared to have faster task completion times for targets with larger ID, yet were insensitive to targets with smaller ID.

Pearson's r characterizes the relationship between the ID and MT. For sighted subjects, the task completion times are relatively close regardless of the distance (correlation coefficient approaches zero) as the target object is located within the reachable area of the hand. Thus, the correlation coefficient for SAR is closer to zero than for the VI and VB, which indicates that SAR is closer to the grasping speed of people with normal vision.

RMSE represents the variation of the MT among trials. It shows that BVI individuals have a relatively stable grasp under SAR feedback (RMSE = 5.970), while sighted subjects have a relatively stable perception in VI group (RMSE = 4.855). The task completion time in NS group is inconsistent for both BVI and signted subjects, resulting in a greater RMSE than the other three groups (BVI: RMSE = 19.240, SIGHTED: RMSE = 19.470).

The box diagram of the positioning error is shown in Fig. 7(a). The positioning error refers to the final coordinate error between the cursor and target after pressing the space bar in Fitts' law test. For BVI individuals, it shows that the positioning error for SAR was reduced by 28% compared with VI, 39% compared with VB, and 46% compared with NS. Meanwhile, the positioning accuracy of SAR for the BVI was 22% higher than for sighted subjects. Additionally, the ranking of the average completion time for the four groups is SAR < VB < VI < NS.

Considering that all four feedback methods require users to re-establish the way of perceiving spatial information, it is necessary to evaluate their learning effect. Figure 7(b) shows the average completion time and error band of all subjects in 30 trials, it is found that the learning effect is not significant in SAR, VI, and VB. However, NS has a more significant learning effect, it shows that task completion time was reduced

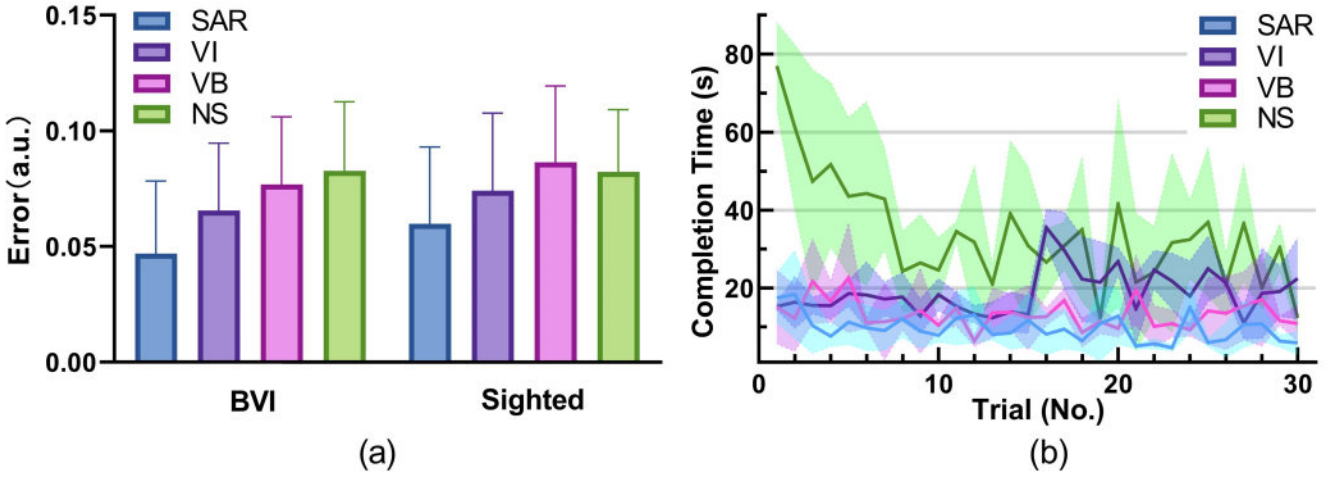

Fig. 7. (a) Positioning error of BVI and sighted subjects in Fitts' law test. (b) Learning effect of four spatial information feedback methods.

after about nine trials. For the results of success rate, SAR, VI, and VB are able to assist the subjects to complete all tasks accurately, with very few failures in which the users were slightly out of the target point. While in the NS group, the majority of subjects experienced long periods of time where they were unable to find the target and eventually had to terminate the trial. In Fitts' law test, we found that NS has significant shortcomings (in ITR, completion time, and RMSE) compared to the other three information feedback methods. As a result, we did not continue to apply it to the desktop manipulate experiment, we will discuss the reasons for the inadequacy of NS in target location tasks in Section V-B.

B. Desktop Manipulate Evaluation

We first evaluated the performance of the assistance system running on a mobile device: The frame rate of the hand and object recognition is 15 fps. The overall CPU usage is 43%, and the memory usage is 400 MB, which is conductive for operations on other mobile devices. Then we evaluated the target positioning accuracy of the Realsense D435 camera. The subjects wore a visual helmet in a normal sitting position. The height of the camera from the desktop was $53cm$, and the viewing angle of the camera covered an inverted trapezoidal area of the desktop (shown in Fig. 8), with an upper base of $40cm$ and a lower base of $70cm$, a height of $45cm$, and a coverage area of $2475cm^2$. We segmented the viewing area into a rectangular grid to measure the errors between the true position of the grid points and the recognition results of the camera after the coordinate system transformation. The results show that the point cloud data for the RGB-D camera provided correct coordinates under the UCS after the pose estimation and coordinate system transformation. The average error for all grid points was $0.43 \pm 0.14\ cm$.

Finally, we evaluated the task completion time and success rate of sighted and BVI subjects in the desktop manipulation experiments, as shown in Fig. 9. The experimental results indicate that the success rate of sighted subjects in completing tasks is slightly greater than that of BVI, and the success rates under the three feedback strategies are similar. However,

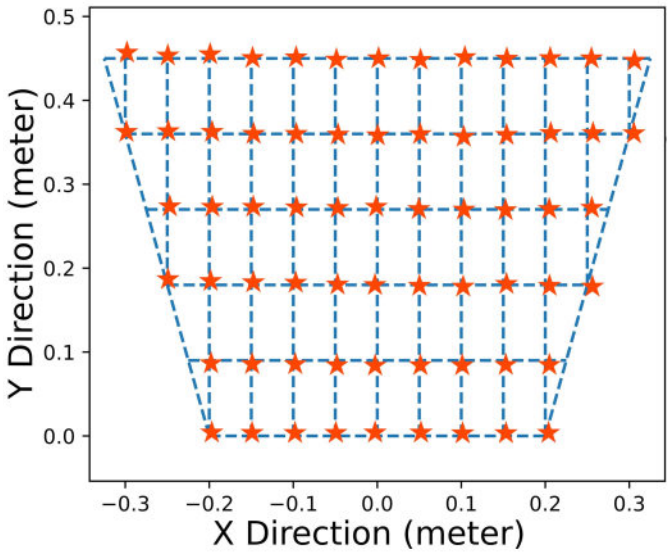

Fig. 8. 3D coordinate positioning accuracy of the RGB-D camera.

compared with the other two solutions, SAR greatly shortens the completion time, which contributes to a smooth user experience.

V. Discussion

A. Analysis of Experimental Results

The purpose of Fitts' law test is to objectively evaluate the performance of spatial information feedback between SAR and three other baseline feedback methods. The experimental results show that SAR improves the ITR for BVI, which is consistent with the results demonstrated by Dunai *et al.* [37] and Frauenberger *et al.* [19].

The desktop manipulation experiment verifies the full target location performance of StereoPilot. Multiple nearby objects placed in the experiment were designed to test the precise localization ability of the system. The experimental results show that the final task success rate for BVI individuals is lower than that for sighted subjects. The main reason is that the recognition accuracy of RGB-D cameras on target objects

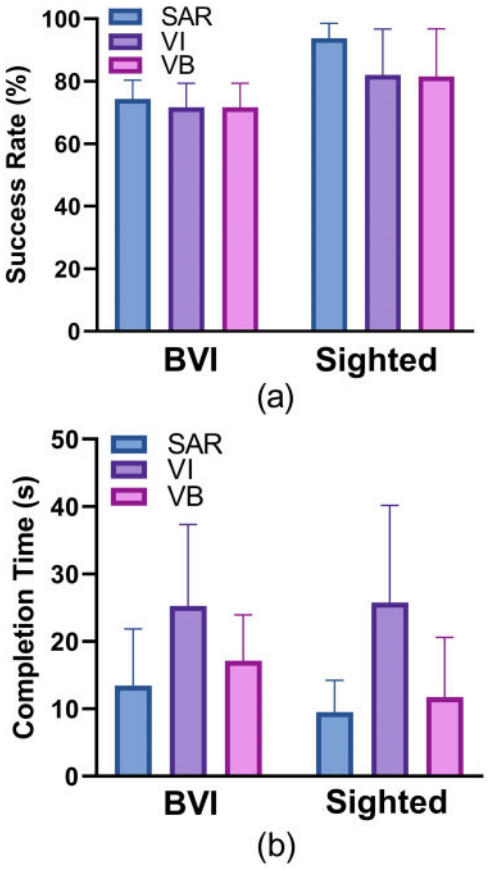

Fig. 9. (a) Success rate and (b) completion time in desktop manipulation experiment.

and hands decreases when the object is blocked by the hand, which can easily cause grasping errors. Sighted subjects can better adapt to low-accuracy spatial information and adjust the perception strategy in time, which might be due to their prior visual experiences.

B. Limitations of vOICe for Target Location

In the Fitts' law test, the NS group performed significantly inferior to the SAR, VI, and VB groups in terms of ITR, completion time, and RMSE, so it's reasonable to believe that the NS group will still yield poor results when CV techniques are incorporated for real-world recognition. Therefore, the NS group was no longer included in the desktop manipulation experiment.

The problems with vOICe in the Fitts' law test are manifold. First, we found that, while vOICe can quickly provide subjects with certain information in the graphical interface, such as the approximate orientation of the two objects (target and cursor) in the interface, most users were unable to distinguish between the cursor and the target. This is because they have a similar shape, size, and grayscale value, resulting in a resemblance in sonification depiction. This perplexity lasted throughout the location task, and users had to listen for sound changes while using the mouse to discern the difference between targets and cursors. Changing the target and cursor's shape and size improves the degree of distinction, but it also raises the chance of the cursor missing the target's center (because when the cursor stays in the target area, it is difficult for the user to further distinguish the exact position of the cursor).

In addition to the confusing identity information between the target and the cursor, when the target and the cursor are close together, the ambiguous temporal phase difference and sound frequency difference perplexed subjects, making it more difficult to understand the correct navigation information expressed by the earcon of vOICe. As a result, when the target and the cursor were far apart, it was simpler for subjects to identify them instantly as the temporal phase difference and frequency difference of the sound were evident, allowing them to complete the target location task faster.

Finally, vOICe's snapshot update rate is only 1 fps, which is insufficient for real-time spatial information feedback. For instance, subjects are more likely to overshoot upon moving the cursor to the target, since they are unable to obtain the updated location information in time. Furthermore, vOICe has a higher learning cost than other feedback modalities, making participants exhausted and frustrated.

C. Comparison Between VB and VI

During Fitts' law test and the desktop manipulation experiment, the VB shows a better performance than the VI. This is because the continuous information feedback characteristic of vibration helps increase the ITR, and users can actively perceive orientation information instead of passively receiving navigation instructions. However, the limited number of vibration units causes the accuracy of the information obtained from VB to be lower than that of SAR while also exhibiting some other unavoidable deficiencies: 1) The vibration module needs to be in close contact with human skin. When used in daily activities, the barrier from clothes or the displacement of units can easily weaken the user's sensory vibration capability; 2) Additionally, the noise of the vibration motors could make users embarrassed in public. Even though more vibration units may contribute to finer orientation perceptions, the trade-off between the wearing convenience and the number of vibration motors gives SAR more abundant spatial information than VB.

D. Future Works

The developed StereoPilot still needs some improvements. 1) Some subjects reported that the rendered spatial audio has some deviations from the actual orientation. This may be due to the incompatibility of the generalized HRTF model on some subjects, which takes users a certain amount of time to adapt. This can be solved by establishing a subject-specific HRTF model and developing a more realistic spatial audio rendering technology. 2) The shortcomings appear in the errors of machine vision recognition in the process of human-environment interactions. Under certain situations, the occlusion of hands and objects may cause positioning errors and result in incorrect spatial information in virtual environments [38]. The robustness of objects and hand recognition needs to be improved in real-time video streams. In addition to the above shortcomings, future work includes developing existing projects onto mainstream mobile platforms (such as Android or iOS) to establish a publicly available software and hardware platform.

VI. Conclusion

This paper develops a wearable target location system to help BVI individuals intuitively perceive and interact with the environment. We first introduced the environmental information perception method based on SAR. The results of Fitts' law test show that SAR can effectively improve the ITR of location information for users. Then we introduced StereoPilot, which integrates a wearable visual perception module and SAR feedback strategy. Based on computer vision technology, this provides environmental information perception and target location for users. We designed a series of experiments to compare SAR with current mainstream auditory and haptic feedback methods. The experimental results on BVI individuals show that users can perceive location information faster by the instincts of sound localization to facilitate their spatial cognition.

References

[1] M. Eimer, "Multisensory integration: How visual experience shapes spatial perception," *Current Biol.*, vol. 14, no. 3, pp. 115–117, 2004.

[2] H. L. Pick, D. H. Warren, and J. C. Hay, "Sensory conflict in judgments of spatial direction," *Perception Psychophys.*, vol. 6, no. 4, pp. 203–205, Jul. 1969.

[3] R. Bourne, "Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the global burden of disease study," *Lancet Global Health*, vol. 9, no. 2, pp. e130–e143, Feb. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214109X20304253

[4] Y. H.-L. Luo and L. da Cruz, "The Argus II retinal prosthesis system," *Prog. Retinal Eye Res.*, vol. 50, pp. 89–107, Jan. 2016.

[5] D. K. McGookin and S. A. Brewster, "Understanding concurrent earcons," *ACM Trans. Appl. Perception*, vol. 1, no. 2, pp. 130–155, Oct. 2004, doi: 10.1145/1024083.1024087.

[6] N. A. Giudice and G. E. Legge, "Blind navigation and the role of technology," in *The Engineering Handbook of Smart Technology for Aging, Disability, and Independence*. Hoboken, NJ, USA: Wiley, Jan. 2008, pp. 479–500.

[7] A. Helal, S. E. Moore, and B. Ramachandran, "Drishti: An integrated navigation system for visually impaired and disabled," in *Proc. 5th Int. Symp. Wearable Comput.*, 2001, pp. 149–156. [Online]. Available: http://ieeexplore.ieee.org/document/962119/

[8] P. B. L. Meijer, "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 2, pp. 112–121, Feb. 1992.

[9] S. Brewster, "Visualization tools for blind people using multiple modalities," *Disability Rehabil.*, vol. 24, nos. 11–12, pp. 613–621, Jan. 2002, doi: 10.1080/09638280110111388.

[10] A. Crossan and S. Brewster, "Multimodal trajectory playback for teaching shape information and trajectories to visually impaired computer users," *ACM Trans. Accessible Comput.*, vol. 1, no. 2, pp. 1–34, Oct. 2008, doi: 10.1145/1408760.1408766.

[11] D. Ni, L. Wang, Y. Ding, J. Zhang, A. Song, and J. Wu, "The design and implementation of a walking assistant system with vibrotactile indication and voice prompt for the visually impaired," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2013, pp. 2721–2726.

[12] D. Ni, A. Song, L. Tian, X. Xu, and D. Chen, "A walking assistant robotic system for the visually impaired based on computer vision and tactile perception," *Int. J. Social Robot.*, vol. 7, no. 5, pp. 617–628, Nov. 2015.

[13] I. Ulrich and J. Borenstein, "The guidecane-applying mobile robot technologies to assist the visually impaired," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 31, no. 2, pp. 131–136, Mar. 2001.

[14] M. Bonani, R. Oliveira, F. Correia, A. Rodrigues, T. Guerreiro, and A. Paiva, "What my eyes can't see, a robot can show me," in *Proc. 20th Int. ACM SIGACCESS Conf. Comput. Accessibility*, New York, NY, USA, Oct. 2018, pp. 15–27, doi: 10.1145/3234695.3239330.

[15] D. Chen, J. Liu, L. Tian, X. Hu, and A. Song, "Research on the method of displaying the contour features of image to the visually impaired on the touch screen," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2260–2270, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9591238/

[16] E. Sampaio, S. Maris, and P. Bach-y-Rita, "Brain plasticity: 'Visual' acuity of blind persons via the tongue," *Brain Res.*, vol. 908, no. 2, pp. 204–207, Jul. 2001.

[17] M. L. Richardson, T. Lloyd-Esenkaya, K. Petrini, and M. J. Proulx, "Reading with the tongue: Individual differences affect the perception of ambiguous stimuli with the BrainPort," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, Apr. 2020, pp. 1–10.

[18] T. Dingler, J. Lindsay, and B. N. Walker, "Learnabiltiy of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech," in *Proc. 14th Int. Conf. Auditory Display*. Paris, France: International Community for Auditory Display, Jun. 2008, pp. 1–6.

[19] C. Frauenberger and M. Noistering, "3D audio interfaces for the blind," in *Proc. 9th Int. Conf. Auditory Display*. Boston, MA, USA: International Community for Auditory Display, Jul. 2003, pp. 1–4.

[20] J. Wilson, B. N. Walker, J. Lindsay, C. Cambias, and F. Dellaert, "SWAN: System for wearable audio navigation," in *Proc. 11th IEEE Int. Symp. Wearable Comput.*, Oct. 2007, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/4373786/

[21] K. R. May, B. Sobel, J. Wilson, and B. N. Walker, "Auditory displays to facilitate object targeting in 3D space," in *Proc. 25th Int. Conf. Auditory Display (ICAD)*, Jun. 2019, pp. 155–162. [Online]. Available: http://hdl.handle.net/1853/61542

[22] X. Hu, A. Song, H. Zeng, and D. Chen, "Intuitive environmental perception assistance for blind amputees using spatial audio rendering," *IEEE Trans. Med. Robot. Bionics*, vol. 4, no. 1, pp. 274–284, Feb. 2022.

[23] S. Tian, M. Zheng, W. Zou, X. Li, and L. Zhang, "Dynamic crosswalk scene understanding for the visually impaired," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1478–1486, 2021.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034. [Online]. Available: http://ieeexplore.ieee.org/document/7410480/

[25] X. Zhang *et al.*, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*.

[26] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.

[27] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, Feb. 1999, doi: 10.1038/17374.

[28] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 231–249, 2001.

[29] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[30] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2001, pp. 99–102.

[31] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 583–593, Mar. 2018.

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

[33] Z. Wei, A. Song, and X. Hu, "Object localization assistive system based on CV and vibrotactile encoding," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Glasgow, Scotland, Jul. 2022.

[34] D. L. Mansur, M. M. Blattner, and K. I. Joy, "Sound graphs: A numerical data analysis method for the blind," *J. Med. Syst.*, vol. 9, no. 3, pp. 163–174, Jun. 1985.

[35] P. M. Fitts and B. K. Radford, "Information capacity of discrete motor responses under different cognitive sets," *J. Exp. Psychol.*, vol. 71, no. 4, pp. 82–475, Apr. 1966. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/5909070

[36] I. Scott MacKenzie, "Fitts' throughput and the remarkable case of touch-based target selection," in *Proc. Int. Conf. Hum.-Comput. Interact.* Los Angeles, CA, USA: Springer, Aug. 2015, pp. 238–249.

[37] L. Dunai, I. Lengua, G. Peris-Fajarnés, and F. Brusola, "Virtual sound localization by blind people," *Arch. Acoust.*, vol. 40, no. 4, pp. 561–567, Dec. 2015. [Online]. Available: http://journals.pan.pl/dlibra/publication/116685/edition/101416/content

[38] C. Shi, D. Yang, J. Zhao, and H. Liu, "Computer vision-based grasp pattern recognition with application to myoelectric control of dexterous hand prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 2090–2099, Sep. 2020.