AiSee: An Assistive Wearable Device to Support Visually Impaired Grocery Shoppers

ROGER BOLDU, Augmented Human Lab, The University of Auckland, New Zealand
DENYS J.C. MATTHIES, Augmented Human Lab, The University of Auckland, New Zealand
Technical University of Applied Sciences Lübeck, Lübeck, Germany
HAIMO ZHANG, Augmented Human Lab, The University of Auckland, New Zealand
SURANGA NANAYAKKARA, Augmented Human Lab, The University of Auckland, New Zealand

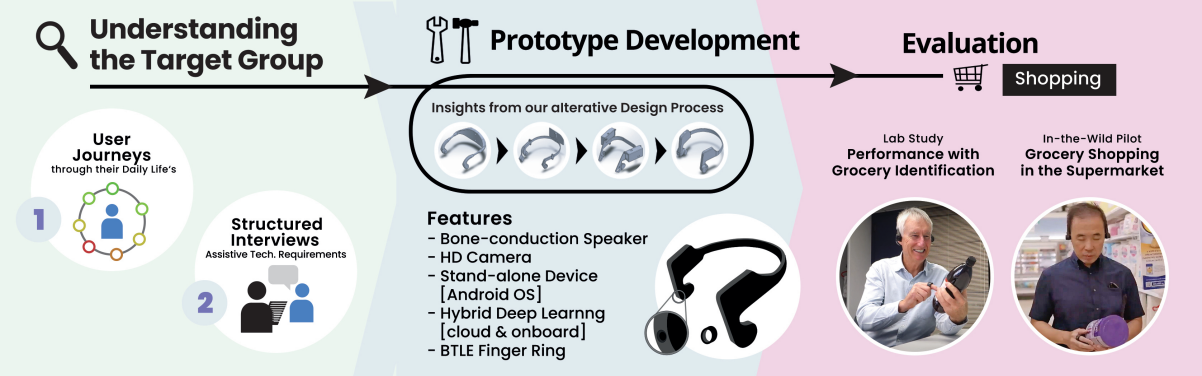

Fig. 1. PVI have developed special habits to cope with daily tasks, such as grocery shopping. We conducted user Journeys walking us through their daily life to understand their habits and challenges, which revealed grocery shopping as a major problem. We also conducted Structured Interviews at which we elicited requirements an assistive device must fulfill. Based on our findings, we iteratively developed a prototype by constantly involving our target group. Finally, we evaluated the performance of our device for the task of grocery item identification as well as in-the-wild pilot study to see how users would actually interact in a supermarket.

People with visual impairments (PVI) experience simple tasks, such as grocery shopping, to be an essential difficulty. Although the recent emergence of AI-technology has been dramatically improving visual recognition capabilities, the application to the daily life of PVI is still complex and erroneous. For example, image recognition engines require a clear shot of the targeted object and a contextual understanding of the information the user requires. In this paper, we aimed to understand the PVI's needs and their pain points in the task of identifying grocery items. Following a user-centered design process, we iteratively

Authors' addresses: Roger Boldu, Augmented Human Lab, The University of Auckland, 70 Symonds Street, Auckland, 1010, New Zealand, rboldu@ahlab.org; Denys J.C. Matthies, Augmented Human Lab, The University of Auckland, Auckland, New Zealand
Technical University of Applied Sciences Lübeck, Lübeck, Germany, denys@ahlab.org; Haimo Zhang, Augmented Human Lab, The University of Auckland, Auckland, New Zealand, haimo@ahlab.org; Suranga Nanayakkara, Augmented Human Lab, The University of Auckland, Auckland, New Zealand, suranga@ahlab.org.

2474-9567/2020/12-ART119 $15.00
https://doi.org/10.1145/3432196

tailored a truly wearable assistive device in the shape of a bone-conduction headset to the needs of our target group. We then evaluated the performance of our prototype, showing a success rate of 80% in recognizing grocery items in a controlled environment. We conclude with a pilot deployment demonstrating how our device can support grocery shopping in-the-wild.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: Finger Touch, Localization, Force, Acoustic

ACM Reference Format:
Roger Boldu, Denys J.C. Matthies, Haimo Zhang, and Suranga Nanayakkara. 2020. AiSee: An Assistive Wearable Device to Support Visually Impaired Grocery Shoppers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 119 (December 2020), 25 pages. https://doi.org/10.1145/3432196

1 INTRODUCTION

People with visual impairments (PVI) face increased challenges, particularly with object identification. This task is crucial for simple and complex decision-making, such as distinguishing groceries or navigation in unknown environment [11, 13]. In these scenarios, relying on assistance from sighted people is often inevitable.

However, for tasks, such as identifying groceries and purchasing them at the supermarket's cashier, assistive technology can significantly help, as shown by recent related works [9, 37, 64, 65]. Meanwhile, existing smartphones are able to empower PVI with assistive apps [34, 43, 52]. These apps become increasingly powerful but yield the drawback of requiring an explicit interaction and creating an unfavorable hands-busy situation. We can overcome this by integrating technology within our bodies [18] and making interaction more seamless [55]. Wearable assistive technology provides a solution and has been previously used to empower users with more natural interaction [47, 49]. Still, these research-based wearable systems are not intended for use outside the laboratory environment and worn all-day. They are often tethered to an external computational device and thus not truly mobile.

In this research, we engage with PVI to better understand their interaction with the environment when it comes to object identification, in particular with groceries. Further, we seek to gain an understanding of their requirements for an everyday wearable assistive device. Based on these insights, we designed and developed an embedded wearable device that aims to provide PVI with constant assistance tailored to their needs. By utilizing a user-centered design process[50], we made several iterations of our design to improve the user experience. Our prototype incorporates a micro-camera that *looks at* the user's field of view, cloud artificial intelligence algorithms that *understands* what the user is pointing at in the captured image, and a bone conduction headphone that *tells* information to the user without blocking the ears (Figure 1 - *Prototype*). Evaluations were conducted with a total of 16 participants to gain insights on the performance, the usability of our prototype, and to test the robustness of our prototype. In summary, our contributions (following Wobbrock et al. [58]) are:

- **artifact**, an assistive bone-conduction headset that adds to the previous body of work on wearable assistive devices for PVI. This includes 1) an understanding of the specific needs via user journey with 6 PVIs and interviews with 5 PVIs; 2) detailed insights on our technical design decisions derived during the user-centered iterative design process.
- **empirical**, an evaluation of our assistive device with a total of 16 different PVIs. This includes 1) pilot study with 6 blind participants; 2) follow up study with 9 participants with a variety of visual impairments; 3) real-world pilot deployment with 1 blind participant.

2 RELATED WORK

Previous work identified and provided solutions for PVI's various needs, including navigation [15], shopping [9, 21, 35–37, 44, 51, 65], and text reading [20, 29, 47]. In addition, there are many commercial solutions that support

navigation [2–4, 27] and reading [16, 19, 57]. However, grocery shopping in unknown environments remains a great challenge for PVI in which they often rely on sighted people [1, 3, 11, 63].

2.1 Shopping Solutions for PVI

A comprehensive study investigating PVI on the difficulties they encountered while taking photographs [11] found that 28% are related to food or beverage items. Szpiro et. al [51] studied how PVI experience difficulties in locating their desired product on the supermarket shelf. The difficulties experienced include identifying the correct product due to similar product shapes, indecipherable labels due to various font types and sizes, and the product's location on the shelf. Based on these identified difficulties, some solutions have emerged. Trinetra [35] is a phone-based system using barcodes and RFID tags to assist PVI in grocery shopping. ShopTalk [44] is a wearable system with a barcode scanner allowing for product search using verbal directions. A smartphone version of these applications was later developed for PVI users [33]. However, the usage of unique tags, such as barcodes or QR codes, presents scalability limitations. Similarly, in certain scenarios, some information presented on the product may not be available in the linked database.

PVI often use different strategies to identify a product, such as holding the product several inches from the eyes, using a magnifying glass, estimating the product content by the shape and size of the box, or taking a photo of the product with a phone to magnify it. Foo et. al [21] developed a grocery shopping assistant that located products and guided PVI to the specific product location. She used computer vision and special hand gloves for tracking, Wiimote for guidance, and 3D audio effects. Lee et. al [37] investigated various methods including speech, non-speech, haptic vibration, a combination of speech and haptic, and a combination of non-speech and haptic, for guiding PVI in acquiring items. Another used technique is crowdsourcing [1, 3, 11]. This strategy requires assistance from a sighted user, namely, when trying to identify a product when grocery shopping, Yuan et al. [64]. While commonly used, limitations such as time, scalability, price, and privacy exist. In developing AiSee, we aimed to prevent limiting it to a specific database or a remote expert.

2.2 Manual Item Identification:

Before detailing how cameras assist with item identification, we will elaborate on current strategies for manual item identification. A common strategy is to rely on tactile perception using only the hands. Generally, the hands are a core physical interaction channel, as proprioception is well pronounced. Even without vision, PVIs can distinguish two-dimensional and three-dimensional objects by holding and rotating them [39]. Gibson et. al calls this exploratory tactile scanning *"active touch"* [22]. This strategy involves micro-motions to explore and measure the object using haptic perception [39, 66]. However, this strategy alone is not suitable for current camera-based item identification. This method usually relies on a single snapshot, in which the PVI has to hold the object still while preventing occlusion.

2.3 Item Identification Smartphones:

Smartphones have shown to empower PVI with camera-based assistive applications, such as KNFB Reader [19], Aipoly [41], Seeing-AI [43] or TapTapSee [5]. Although the recent emergence of AI-technologies has dramatically improved visual recognition capability, it is still difficult to apply to the daily life of those with visual impairments [25]. It is a known issue of the current vision system. The image recognition engine requires a clear shot of the target object; however, this is difficult for visually impaired people. Kacorri [31] tried to make a custom recognition engine to address this problem. Also, some studies like [30, 56] used audio feedback to correct camera aiming for better image capturing. Targeting wanted information, such as text, can be difficult, since it is often inappropriately framed within the captured image. The task difficulty increases dramatically with decreasing vision capabilities [14]. Cutter and Manduchi further investigated different feedback modalities supporting PVI

to orientate the phone correctly [14]. One of their most crucial findings is that in 59% of all provided guidance, the user only required simple orientation correction to sufficiently frame the target. While the phone presents an advantage being commonly used by PVI, limitations exist. Features, such as the touchscreen and 'voice-over,' are hard to operate [60] and generates a hands-busy situation. Overall, it is difficult to control the phone while in motion.

2.4 Wearable Cameras:

Alternative approaches have emerged, such as mounting cameras onto the user's body by using wearable technology. This approach frees PVI's hands, which is an important perception channel and simplifies the use of technology. Therefore, different parts of the body have been explored for camera placement. For instance, FingerReader 2.0 [9] places a camera on the index finger to provide assistance in shopping scenarios. Similarly, HandSight and FingerReader [46, 49] explored the design space of finger-based text scanning.

Stearns et al. [49] developed and evaluated a prototype exploring how to continuously guide PVI's finger across text using different feedback conditions for a better user experience. Although wearing a camera on the finger may feel natural, object occlusion and image framing were identified as critical issues. Another very popular position to place cameras is the head, namely by incorporating it into a pair of glasses [48, 57, 61]. OrCam [57], a commercially-available device, uses a small camera attached to the frame of a pair of glasses. The device can read text, describe objects, and assist with facial recognition. Another device, Horus [17], is a system that consists of a headset form factor with a camera connected to a waist minicomputer. It uses deep learning algorithms to recognize objects that a user is looking at.

AiSee adds to the previous body of work on wearable assistive devices. By following a human-centered design process, we found reasons to question the typical approach of using glasses augmented with a camera. PVIs often mentioned in interviews experiencing stigmatization when wearing glasses. Therefore, we are proposing an alternative form factor incorporating a discreet bone conduction headphone.

3 UNDERSTANDING THE TARGET GROUP

To understand our target group, we conducted *User Journeys (Subsection 3.1)*, in which PVIs walked us through their daily routines, revealing their habits and challenges. We followed this with a subset of *Structured Interviews (Subsection 3.2)* that aimed to elicit requirements an assistive device needs for PVI to consider it usable. This section documents our process and the resulting findings.

3.1 User Journeys - Daily Routine

A common design research tool used to identify the daily challenges of a certain user group is User Journeys. We conducted a user journey with 6 PVI, (2 females and 4 males) aged between 21 and 72 years (M = 54.1; SD = 18.08). All participants were randomly selected from a list of PVI provided by the local association of visually impaired. Three participants were completely blind, two had retinal pigmentation, and one had diabetic retinopathy.

The interviews lasted for about 1 hour. We asked participants to imagine and describe their day upon waking until sleep with as much detail as possible. The researcher took notes of the overall user journey. Figure 2 compiles the user journey of one of the participants.

From the user journeys, we observed that most participants had a relatively active life. They completed multiple activities, such as going to work, classes, visiting friends, etc. However, we observed that participants P1 (95% vision loss, Male, 61 y/o) and P4 (Completely Blind, male, 72 y/o) had a slightly less active life compared to the others. They stated they usually remained at home all day. We hypothesized that their family environment and personality have a large effect on their daily routines. The participants we accessed and interviewed are those

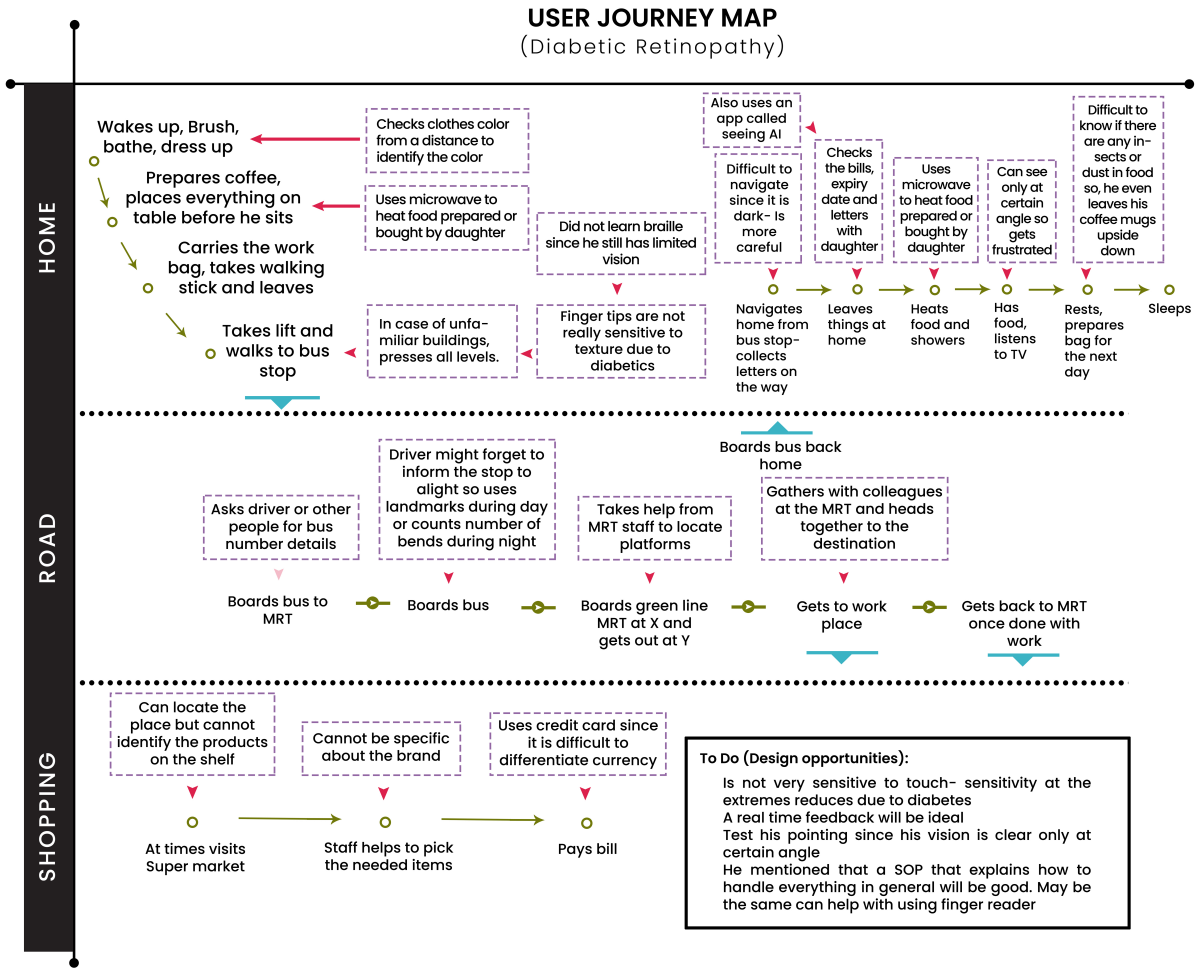

Fig. 2. Example user journey of P5, male of 58 years old with diabetic retinopathy

classed as early adaptors, which generally are better integrated into society. Nevertheless, the vast majority that remains at home could potentially benefit even more from assistive technology.

All 6 participants reported facing difficulties with printed text, such as reading documents, letters, and the news. The other four active participants also reported difficulties with fashion, namely identifying colors to ensure their clothing matched. Except for P2 (Completely blind, Male, 21 y/o), who indicated that he uses private taxi or family transportation, the remaining participants used public transport to get to work, school, etc., and face difficulties mainly with stopping the vehicle (to get in and out).

All 6 participants reported facing multiple challenges with food, such as cooking, grocery shopping, and knowing whether food is in good condition (i.e., not expired). The four most active participants reported going grocery shopping and stated they faced difficulties in locating the products (mainly if they changed their location, or it was an unknown environment), and identifying the details of the product, such as the brand, ingredients, quantity, the expiration date, etc. While grocery shopping, participants reported being dependent on sighted people. Two participants mentioned they immediately sought assistance from supermarket employees to help them obtain specific items. Others reported going with a friend or asking strangers around the store to help them

identify the different products. However, on certain occasions, participants preferred to not request assistance for personal reasons or the nature of the purchase. In that situation, they face difficulties with identifying a specific item they are looking for precisely. Sometimes they rely on guesses as to whether they are purchasing the right product. This problem increases when there is a package similarity between items. For example, brand differentiation with toothpaste may be difficult, given most tubes have similar shapes.

3.2 Structured Interviews - Requirements of an Assistive Device

Having identified some pain points presented by PVI, we wanted to learn how would they envision the requirements of an assistive device that fits their needs.

We recruited 5 participants, where one was blind, while the other 4 had low vision. The participants were randomly recruited through the local PVI community at *(Anonymous city)*. All participants had a full-time job. All participants also had experience with current state-of-the-art assistive devices and were aware of on-trend products in assistive technology field, such as KNFBReader [19], Orcam [57] or Seeing-Ai [43].

This study was designed to be a structured interview. We conducted concise questions in which participants were to answer with dis/agree or provide a rating on a 5-point Likert scale. Our questions regarded system functionality, form factor, camera location, system characteristics, and system interaction.

3.2.1 Daily Use of Existing Wearables: All participants were asked to rate how often they wore devices in 4 different locations, the head, wrist, chest, and finger. The ranking was based on a 5-point Likert scale (1: Never 5 Daily). Participants stated they used smartwatches, such as the Apple Watch, on a daily base (M = 5; SD = 0). 4 out of 5 participants reported wearing headset devices, such as headphones often (M = 4.8; SD = .45). Rings (M = 2.6; SD = 2.19) and chest-worn devices (M = 2.2; SD = 1.64) were worn on rare occasions.

3.2.2 Preference in Wearing Position: All 5 participants were asked to rate their preferred wearing location of a camera-based assistive device. Four positions were to be ranked: head, finger, chest, and wrist. The ranking was done by organizing the four different positions from most preferred (1) to least preferred (3). The position of the head scored the most preferred 1 (M = 1.8; SD = 0.83), followed by the wrist (M = 2.4; SD = 1.1), the finger (M = 2.8; SD = 1.6), and the chest (M = 3; SD = 0.7). Only the position of the head scored a slightly positive result. We suspect that a reason for this result is that participants previously reported wearing head wearable devices more frequently, such as glasses and headphones. To clarify, we asked questions regarding their experience with these further.

- *Glasses:* Only one of the 5 participants reported wearing glasses on a daily basis (M = 1.8; SD = 1.78). Other participants mentioned that glasses often produce awkward situations. Another participant, P4, said: *"When I wear glasses, people expect me to have a certain eyesight, and that could be negative"*. P2 mentioned that glasses obstruct his residual peripheral sight.
- *Headphones:* All 5 participants reported using headphones on a daily basis(M = 4.4; SD = 1.34). Participants reported using standard off-the-shelf headphones. Four participants reported that they own and use a specific type of headphones, bone conduction. Interestingly, bone-conduction headphones are not specifically produced to assist PVI. Yet, all 5 participants perceived it as an assistive tool. They mentioned the usefulness of bone-conduction headphones due to the non-obtrusive audio channel. They suggested using them while doing outdoor activities, such as navigating with Google Maps. One of the participants criticized the quality of audio, while another mentioned that it was slightly awkward for public use.

Finally, the participants were asked to indicate their preference between bone-conduction headphones and glasses to integrate an assistive device. Interestingly, 4 participants, namely those that owned bone-conduction headphones, stated their preference for them. Two participants indicated additional assistive features, such as calling someone by pressing a button directly at the headphones.

3.2.3 Envisioned System Capabilities: Although novel technology can provide powerful capabilities, there are other important considerations, such as "Ease-of-use", "Discreetness", and "Wearing Comfort." The participants were asked to rank these properties and the "Accuracy" based on their importance. Accuracy (M = 3.8; SD = .44), followed by the ease of use (M = 2.8; SD = .44) received higher scores. However, open-ended conversations clearly indicated the importance of "Discreetness".

4 PROTOTYPE DEVELOPMENT

In this section, we provide the technical details of our first iteration of the device. After developing the first version, we took an iterative design approach with multiple PVI to improve its performance and usability. More details can be found in *Section 5*.

4.1 Design Requirements

Informed by the previous interviews and related work [9, 25, 31, 47], we defined a subset of design requirements that we followed to design AiSee.

Form Factor: A device that does not interfere with the daily life activities of the user (i.e. "wear and forget"). However, it needs to be available in a discreet way. Therefore, the device should be wearable and hands-free, enabling the user to use it while performing other activities, such as talking or walking while carrying other objects, such as the white cane. The apparatus should also not diminish any residual gaze the users may have. Also, in case the user is wearing glasses, it should be able to adapt.

Camera Location: The location of the camera should enable the sensor to capture the proximity interaction area of the user without the need for any implicit action. It is also crucial that AiSee is socially acceptable. Without such acceptance, some users reported they would not use it. Therefore, AiSee camera location should be slightly hidden so that others do not target the users as a handicap. However, at the same time, the camera should capture the ROI (region of interest) of the product, while keeping a low level of occlusion. Finally, the camera should have a static location and not shake to extract precise and sharp images.

Processing: AiSee should be able to provide an audio description of any presented item that the user wants to identify. Consequently, the recognition should be based on the features of the item, such as text, color, or shape, rather than training a specific subset of objects. Image processing should be done as fast as possible and in real-time, if possible. The feedback provided to the user should contain enough information for the user to make the right decision about the specific item, while not overwhelming them. Finally, if possible, the processing and power consumption should be low for the device to last a full day with one charge.

Usability: Following the interview insights on the system capabilities, AiSee must enable the users to access the information in a fast and discreet manner. With a natural and straightforward interaction, the user should be able to extract any desired information. The device should be easy to use and discreet to the others.

4.2 Overview System Architecture

The overall interaction flow of the system is a similar approach used in image recognition systems such as Phone Applications [19, 43] or wearable devices such as OrCam [57] (1) The system, by default, is on standby mode to keep low power consumption and waiting for the user to (2) trigger the recognition system. Once the system is triggered, an (3) image is captured. Then, an AI algorithm (4) extracts the region of interest, which is defined by where the user is pointing at. Finally, (5) features from the selected area are extracted, elaborated on a sentence, and (6) delivered through audio to the user.

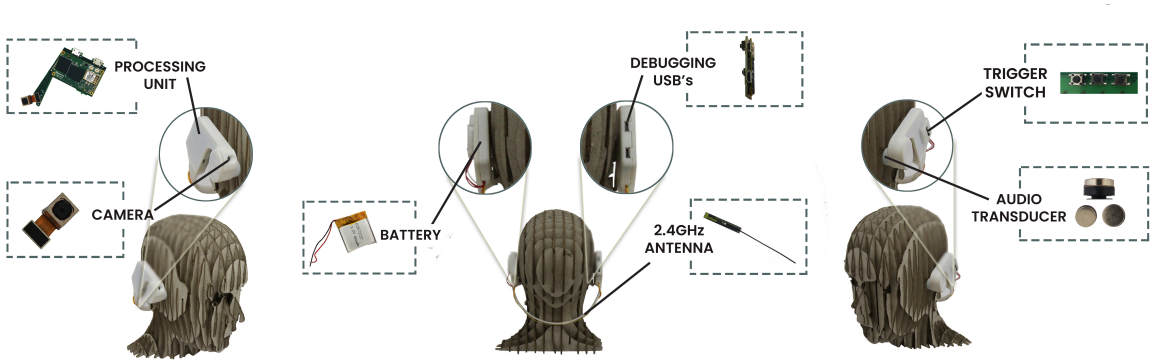

Fig. 3. Overview of the first iteration of the AiSee wearable apparatus. The image shows all the different parts that were integrated inside the headset and their location

4.3 Hardware Overview

The AiSee prototype incorporates three key hardware components: (1) a camera, (2) a processing unit, and (3) 2 bone-conduction transducers.

The processing units consist of an ARM mid-range SoC (Qualcomm Snapdragon 410), quad-core Cortex-A53 with a frequency of 1.2GHz, and 2GB of ram[1]. The camera model selected is a Sony IMX135, which has 13 megapixels of resolution, and the lens provides a field of view of 79.8° with autofocus capability. MIPI CSI interface is used to communicate with the camera sensor. The prototype embeds a WiFi and a Bluetooth module and connects to the cloud through WiFi. A small battery of ~ 400mA is used to power the system. The electronics are placed on the right side of the headset, while the battery is located on the left side (see Figure 4).

With this small and light battery, 4 hours of battery with only a single charge is possible. AiSee power consumption strategy is on the software side. The device is on standby mode over 90% of the time, waiting for the user to trigger recognition. However, when the device is interfacing with the camera and pre-processing the images, we observe significant peaks of current (~ 500mA - 1A) for periods of 2-3 seconds. Also, the highest and most significant power consumption spikes are found when the system is booting the OS (Android), where we observe current peaks of up too 2 A. Although the electronics of the system are designed to handle these currents, we observed some complexity in heat dissipation management. During regular usage, the users did not report facing any problems with heat. However, some users reported being slightly uncomfortable during the booting process of the system. Thus, we slightly improved this problem by relocating the electronics and placing the SoC (which is the one that produces more heat) further from the users' skin. Future versions should have the casing developed from conductive materials such as aluminum for a passive dissipation of heat.

4.4 Form Factor Development

During the design of AiSee, our primary motivations were to achieve a form factor that did not interfere when the user performed other activities and prevent others from targeting the user as handicapped. During the interviews, we found that users were using standard bone-conduction headphones, which are commonly found in consumer electronics (e.g. Aftershokz[2] to interact with phone applications and perform tasks, such as GPS navigation. Following this trend, AiSee adopts this form factor and embeds a small camera to perform visual recognition tasks. The three main design requirements related to the form factor of AiSee were: (1) compact integration

[1] www.varacite.com
[2] www.aftershokz.com/

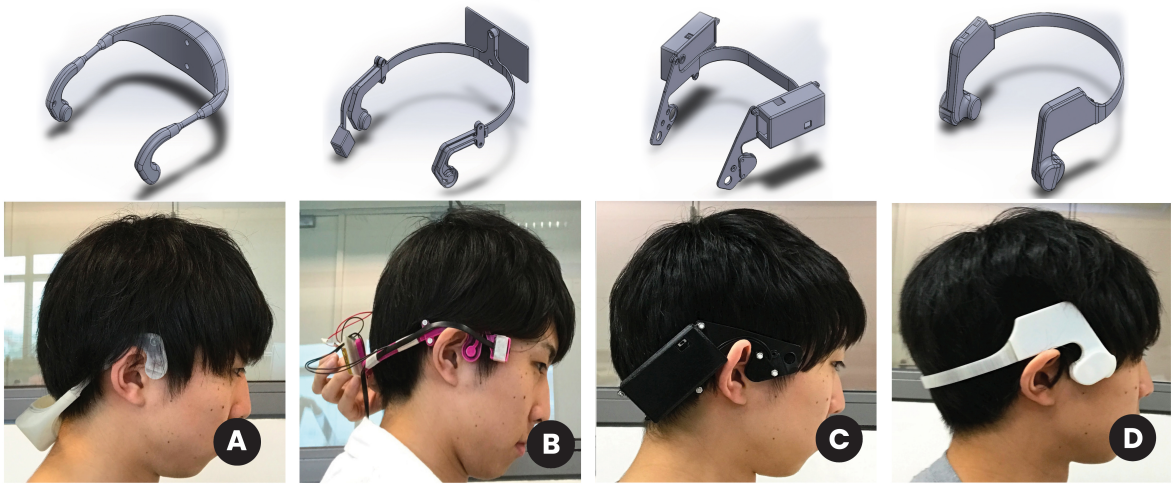

Fig. 4. Different iterations of the device form factor, *(A)* - Functional prototype, which housed all the electronic components, but was fragile and bulky. *(B)* - Modular prototype, which enabled testing of the tightness of bone-conduction headphones and adjustment of the camera location and orientation. *(C)* - Ergonomic prototype, where the components were re-positioned to move the center of mass between the ears. The user still felt uncomfortable for prolonged usage. *(D)* - Compact prototype, where the size and weight of the Headset were reduced with a smaller PCB design. The material of the Headset was changed to increase flexibility, audio performance, and comfort.

of hardware; (2) appropriate tightness of the pieces and pressure to ensure audio quality of bone conduction without being too tight and uncomfortable; (3) camera-mounting that provides a right field of vision without being unobtrusive. Figure 4 shows the form factor iterations and the details appear on each of the caption.

To get the right amount of pressure and a comfortable form factor, we printed the parts with different patterns using 3D printers. The two main casings for the electronics and batteries were printed using a multi-material Object 500 3D printer. While the band that goes around the head was 3D printed with a filament printer, a MarkerBoard was used following horizontal patterns. This approach enabled us to adjust the pressure that was applied to the users. Furthermore, we also used a hot gun to thermoform some parts of the prototypes to create better fits for the users. We developed and assembled a total of 10 of these prototypes, which we later used to evaluate the performance of the device.

4.5 Camera Properties (Angle & Placement)

The two primary considerations for camera placement were the discreetness of the device (not noticeable by others) and the field of view of the camera (not occluded by the wearer herself/himself). Using prototype *(B)* (Figure 4B) and its modular feature, we evaluated the relation between camera angle (5°, 10°, 15°, 20° and 25°), target distance (200 mm, 400mm, 600 mm, and 800 mm), and camera view. As shown in Figure 5, a researcher wore the prototype and pointed at a mark (X shape, 40mm x 40mm) on the wall (height 150mm) with the right hand. This action was repeated for each of the 20 combinations of angle and distance. The views from the camera were recorded and analyzed.

It can be observed that 15 to 20 degrees are more suitable for shorter distances, and 5 to 10 degrees for longer distances. For our system, a 15-degree angle was selected, since the finger is closest to the center of the camera view for shorter distances.

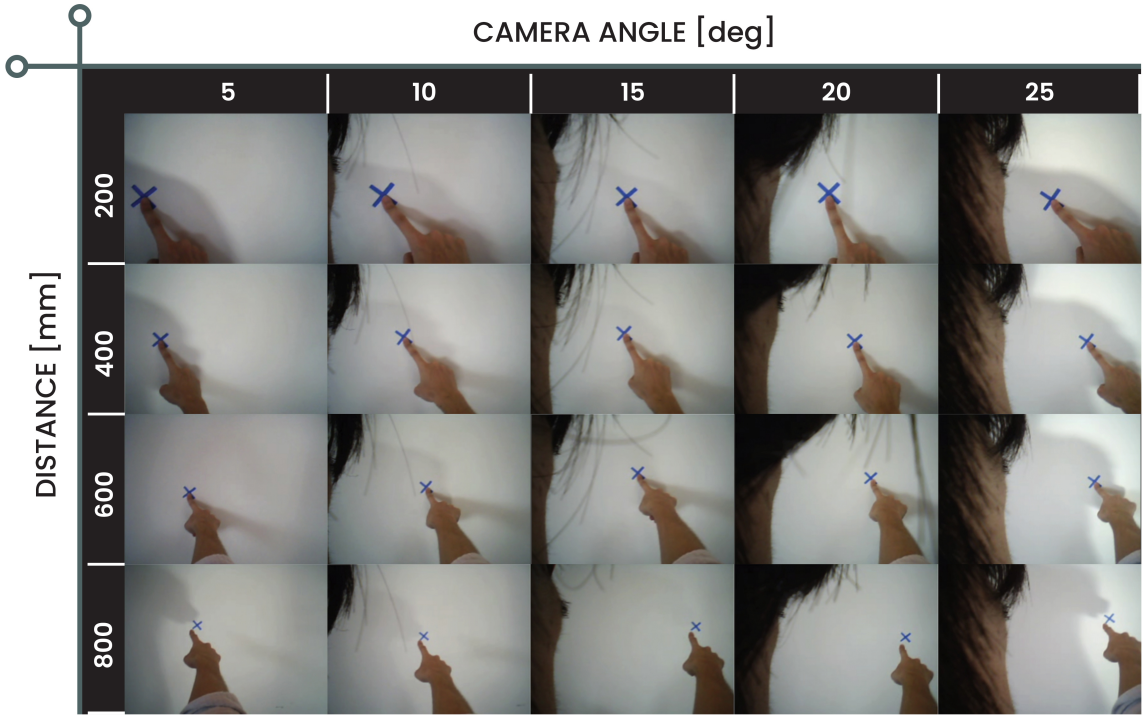

Fig. 5. The table shows the correlation between the angle of the camera and occlusion. The camera should be placed on a discreet location, while still having excellent visibility of the target.

4.6 Bone Conduction Transducers

The selection of bone conduction transducer has direct implications on the audio quality. During the design of AiSee, multiple transducers were considered and evaluated. The main characteristics considered were the price, the sound quality, and the form factor. The final design of AiSee incorporated two transducers BM230, set up on a stereo configuration. We used the existing internal power amplifier of our selected SoC. This uses a classic AB audio amplifier that can deliver up to 3W on transducers of 8 Ohms.

During our initial iterations, we used high-end transducers (ppu > USD 100). However, the users reported that the audio quality was poor. They had problems with hearing the audio clearly. We observed that the main problem was in the assembly methodology, where we attempted to embed the transducers in the casing and hold them by mechanical properties. This approach made the sound present a high distortion (the users described as "metallic sound."). After some iterations and evaluation, we improved the quality by simply gluing the transducer to the casing, reducing the inner casing space, and tightening the enclosure by .5 mm. This solution substantially improved the quality of the audio and enabled us to use USD 3 transducers. When the transducers worked at the maximum level ($0.5W/88dB = -3DB$), AiSee generated a ticklish sensation, which some users liked, but others found it uncomfortable. At the same time, it reduces the privacy of the system, since those in the user's vicinity may be able to hear some audio.

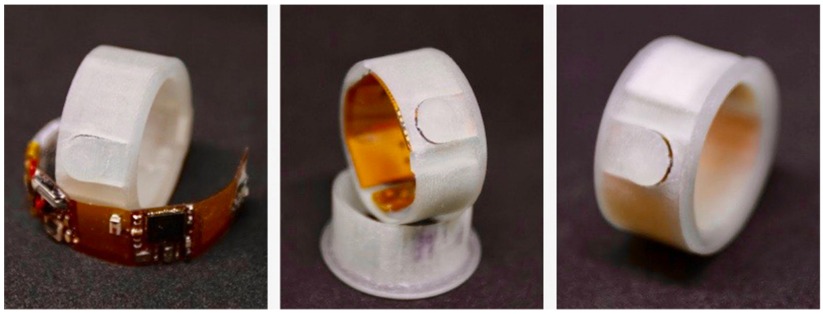

Fig. 6. Bluetooth wearable ring used to trigger the headset. The ring contains a custom-made FPCB with a micro switch and a Bluetooth Low Energy IC (Nrf51822)

4.7 Software Design

The software system adopts a hybrid approach to maximize battery life and flexibility in deployment. The embedded system runs Android (Lollipop 5.1.1) as OS. The system connects to a peripheral board that integrates three different switches to trigger a recognition event (see Figure3). When a recognition event is triggered, an image is captured and pushed to the cloud. On the cloud, a Java-based server analyzes the image by using two algorithms in series. First, it extracts the object's region of interest (see section *5.2*), and second, it extracts the features of this one. Second, we use Google Cloud Vision API [6, 54] to extract further features (see section *5.2*). As output, the system elicits textual descriptions of the scene in three main categories logos, text, and general object characteristics (e.g., book, electronics, etc.) and returns the information in plain text to the system. Finally, the generated answer is sent back to the embedded device, which organizes and filters the information (see section *5.2*), before converting this one into audio with the Google text-to-speech engine.

5 PROTOTYPE ITERATION

Once we had an initial version of the device (see Figure 3), we performed an iterative design approach [50] to identify potential problems and improve its usability. The iterations were done throughout multiple evaluations with a subset of 6 PVI aged between 22 and 61 years (M = 45.83 SD = 19.75; 5 male and one female) from the local association of PVI at *anonymous city* The participants participated in one or multiple one-on-one interviews, which helped evaluating and providing feedback about the device over a total period of 6 months. Not all participants were part of every iteration. Also, one of the participants became a member of our research institution for a period of 3 months and provided extensive feedback.

5.1 Iteration 1 - Fast and Easy Interaction

The first version of AiSee contained a small subset of triggers on the left side of the headset, which the user needs to press to trigger the recognition. After evaluating it with two participants, we quickly observed that the interaction seemed slightly confusing for the users. They had to hold the item in one hand while pressing the button with the other hand. Even though the interaction itself was hands-free, the user had to perform an explicit action, namely raising their hand to the headset and press the button.

To keep the interaction as unobtrusive and discreet as possible, we embedded a subtle switch on a wearable ring (see Figure 6) worn at the index that would trigger the recognition process. Simply pressing the switch using the thumb is a minimal microgesture [59] that potentially enables a reflexive interaction [42] and thus seems to be an optimal input technique. This thumb-to-index gesture can be considered as hands-free and may be performed with little effort in parallel to other real-world tasks [10]. The device consisted of a custom-made

Fig. 7. Bottom row: original images extracted from the camera after identifying the gesture. Top row: croped images after extracting the Region of interest.

FPCB (Flexible printed circuit), a micro switch, and a Bluetooth Low Energy IC (Nrf51822), that communicates with the headset (see Figure 6). The device used a 10mA curved battery to power the system. The battery lasts for up to 30 days when fully charged. The ring charges using an external apparatus, a ring casing. The casing was 3D printed in two parts by an Object500 printer. The Bluetooth communication between both devices is based on the event delivery of a status change. We did face some complications with the communication between the ring and the headset due to reduced signal transmission. We improved the communication by re-arranging the 2.4GHz headset antenna-FXP831 (see Figure 3) to the front of the headset instead of the back.

5.2 Iteration 2 - Information delivery - Software AI

AiSee relies on AI algorithms to extract as many features as possible from the captured image and to generate a sentence that can then be delivered through audio to the user. In the beginning, our goal was to provide as much information to the user as possible, with the hypothesis that he or she would extract the relevant information and thus make the right decision. Therefore, to extract as many features as possible, and to make the system less specific, we decided to use a state-of-the-art general propose AI cloud service API[3].

Region of Interest: To reduce the unrelated information and center the recognition around the item, we decided to crop the image and remove the areas without the object before sending it to the AI engine. This (i) reduces the chances of identifying irrelevant features from the surroundings, and (ii) extracts more information from the object the user is targeting. The gesture of point at the bottom left corner of the item was introduced to indicate to the system the ROI (see Figure 7 bottom row). A customized Single-Shot Detector (SSD) [38] model was trained to detected the hand pointing and extracted its coordinates. The coordinates were later used to crop the image vertically and horizontally, starting from the index finger (see Figure 7 top row). Finally, the cropped image was then sent to the AI engine. The SSD model was done by labeling over 100 pointing images [32, 53]. We used TensorFlow 1.21 object detection API [23].

[3]https://cloud.google.com/vision

Information Filtering: The used AI service extracted different features from the image (see Figure 8), related to the text, colors, Logos, object classification, web search, and others. However, while evaluating the device with participants, we observed that providing excessive information that is not 100% relevant for the recognition (e.g., junk food, drink, etc.), produced confusion, and made the users lose their concentration on the outcome. While the system provided the user with the exact information required in some locations, the user could not extract it. Instead, he or she remembered the higher class results, which are less useful for the recognition. Based on these observations and related work [9], we decided to (1)organize the information to be more relevant for the users. From our observations, this includes logos, text, and categories. Then, we also (2)implemented word filters to remove layers of information from the categories section that produced confusion to the user, such as *"junk food," "drink," "can," "finger or arm," etc.* As a result, we were able to reduce the loop time, and the users seemed to improve their performance. However, we envision that future work should focus on understanding the overall problems faced in these scenarios.

5.3 Iteration 3 - Sound UX Notification

During the evaluation, participants suggested that a maximum number of reasonable attempts they will perform to recognize an item are 3. *"In maximum three tries, the application needs to accomplish its goal, e.g., identifying the product. Otherwise, it leads to frustration."*. Also, we observed a trend where users preferred to perform multiple quick iterations that contained less information rather than slow iterations that contained lots of information. I.e., if it is slow and the user pointed on the wrong side of the object, that will lead to frustration.

Also, we observed that when the image contained a large amount of information, such as a document or the ingredients of a food product, the processing time of the image was long. The slowest interaction we faced was when a PVI attempted to read a full printed document. Due to the amount of information contained in the image, the system took about 5.2 seconds to generate an answer. Although the participant understood the system could take longer to provide the image, the lack of feedback generated uncertainty. The user kept triggering the device without waiting for a response. To improve the usability of the system, we decided to add notification sounds. The notification sound is played when (0) the system is booting (guitar sounds), (1) when the image is captured (camera shutter sound effect), (2) when the image is being processed (ultrasound radar sound effect), and (3) when the result is received (notification sound effect). The sounds were extracted from a UX library. We received positive feedback from users when adding this essential improvement for the user experience.

6 EVALUATION

After a long iterative process, we decided to perform a user evaluation aiming to understand the performance and usability of the apparatus.

Table 1. Participants involved during the pilot Evaluation. Demographic information, percentage of items identified correctly (%Success Rate), the average average number of shots per task

Case	Age	Gender	Visual Acuity	WHO *Score*	%Success Rate	(M) N-shots	SD
P1	31	Male	Light	Blindness (cat=5)	75%	3.37	1.61
P2	25	Male	No Light	Blindness (cat=6)	69%	3.05	1.59
P3	42	Male	No Light	Blindness (cat=6)	80%	2.8	1.56
P4	39	Male	Light	Blindness (cat=5)	87%	3.17	1.61
P5	33	Female	Light	Blindness (cat=5)	79%	3.10	1.56
P6	32	Male	No Light	Blindness (cat=6)	91%	3.17	1.58

6.1 Pilot Study

We initially began with a pilot study to get some first impressions on the device performance, task performance, and usability problems. We invited 6 participants (5 male, 1 female) aged between 25 and 42 years M = 33.66 (SD = 6.05) individually for five sessions lasting an hour each. All 6 participants were completely blind (cat. 5 & 6 [7]), with only 3 participants capable of spotting lights and bright spots. However, the cause of blindness differed between participants, such as Birth or Uveitis (see Table 1).

Motivated by our previous user journeys, we defined the following tasks. The user was asked to differentiate between similar items (i.e. the researcher asked questions such as: *"Is this a can of coca-cola, red-bull, or sprite?"*). In each session, we asked the participant to perform 27 recognition tasks with 9 new products (types: 3 boxes, 3 cans, 3 bags) × 3 repetitions. The grocery items were given in a random order and orientation. By the end of the 5 sessions, each participant had performed 135 recognition tasks with a total of 45 different items. The success rate was calculated as the percentage of correct answers. Furthermore, we collected the number of shots (images taken) the participant performed before providing the answer. Participants were allowed to trigger the device multiple times, but could only provide one answer per item. Also, the researcher did not inform the participants if the provided answer was correct or incorrect. Additionally, at the end of sessions 1, 3, and 5, participants completed a standardized task load questionnaire (NASA-TLX [26]) and a standardized usability questionnaire (SUS [12]). One participant missed session 5 due to personal reasons.

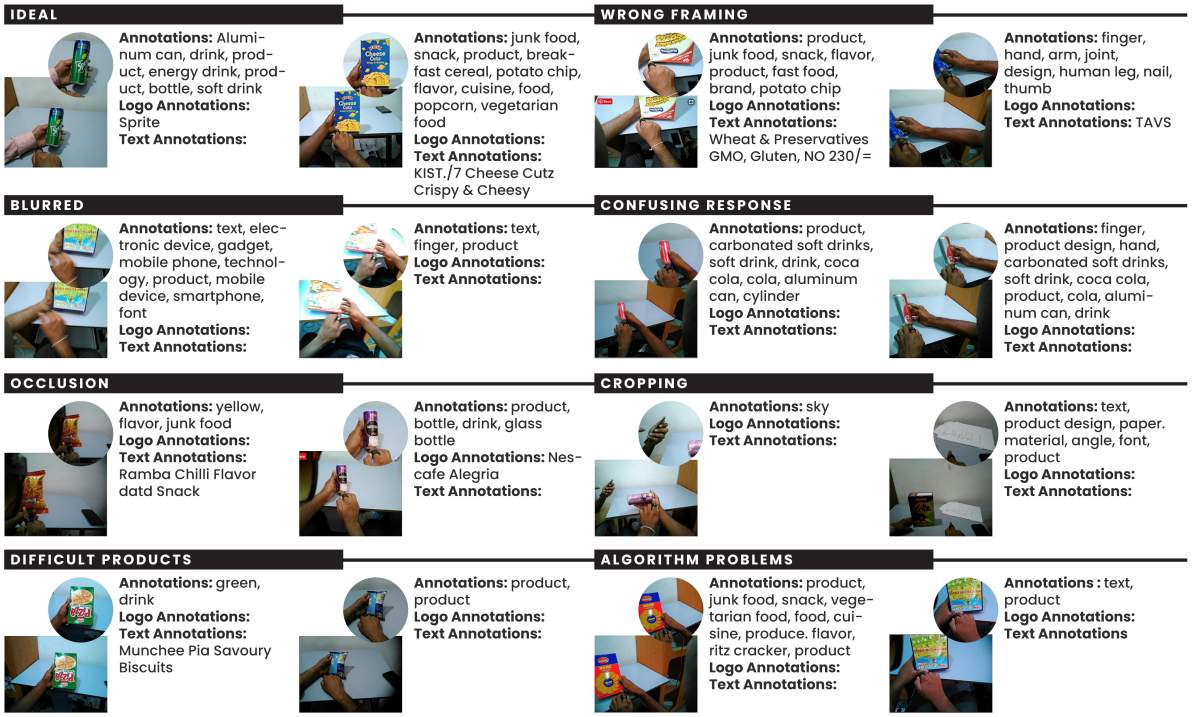

Fig. 8. Subset of images that show the different challenges faced to provide reliable feedback. Each of the instances contains its original picture and the ROI picture. Also, at the bottom of each image, there is the answer given by the algorithm before applying a filter.

6.1.1 User Performance. The overall success rate of the presented task was 80%, with an average number of shots of M = 3.1 (SD = 1.58). We observed a correlation between success rate and object shape. Boxes (86% success) had the highest success rate, compared to bags (79%) and cans (76%). Throughout the study, we saw that inaccuracies are rooted in the targeting style, particularly when a participant could not estimate where potential information was located on the object, such as with cans. Further evaluations should explore a larger variety of items, with a more generic and complex task. We did not observe a learning effect throughout the five sessions in terms of performance, suggesting that it is simple enough to use without much practice.

6.1.2 Task Load. The following average scores were gathered from the pilot study: Effort: M = 15.08 (SD = 7.75), Mental Demand: M = 11.45 (SD = 10.5), Temporal Demand: M = 9.88 (SD = 7.47), Performance: M = 9.04 (SD = 6.42), Frustration: M = 4.79 (SD = 6.62), Physical Demand: M = 3.74 (SD = 5.83). These scores represent the break down of the weighted TLX score. i.e. ($X_Rating \times X_Weight \div \sum Weight$).

Overall average TLX score was ~ 54. The task load seems consistent throughout multiple sessions. The results of the NASA TLX inform that level of Frustration, and Physical Demand were relatively low compared to Effort, Mental Demand, and Temporal Demand. These results should be taken into consideration for future usability improvements.

6.1.3 Usability. We observed an increment on the SUS scores as the sessions progressed, indicating that users became familiar with the device. Session 1: M = 52.5 (SD = 15.08), session 3: M = 67.08 (SD = 6.78), and session 5: M = 68.75 (SD = 10.89). Following Brooke [12], our system can be classified as usable after the participant underwent training for five sessions.

6.1.4 Image Quality: The collected images were organized based on the Levenshtein distance [28] between the algorithm answer and the list of original products. A researcher then manually annotated a subset of 150 images from the list (50 top, 50 bottom, 50 middle). Figure 8, which summarizes the different image quality challenges.

6.2 User Study

Informed by the pilot study, we proceeded to evaluate AiSee. Based on our results we altered the study design as follows:

(1) Single session. For practical considerations regarding the length of the user study and the absence of a learning effect, we decided to shorten the evaluation to a single session.
(2) Participants with different eye conditions. The pilot study had participants with homogeneous eyesight (totally blind). This study considers low vision individuals and explores how they would benefit from AiSee.
(3) A greater variety of items, especially round shape items, which the pilot study demonstrated had lower accuracy.
(4) Instead of a multiple choice task used in the pilot study, participants had to describe the item and volume and/or weight of it.

6.2.1 Study Design.

Apparatus: The apparatus used is the same as described in the iterative design process. However, to extract more insights and speed up the user evaluations, we developed a mobile phone application enabling the researcher to control the system during the evaluation. With this tool, the researcher, at any given time, can set-up the WiFi network of the device, adjust the audio volume, and adjust the speed of the text-to-speech engine. Additionally, the app can visualize what the camera is seeing and visualize the results extracted from the image processing algorithm. Finally, if needed, the researcher can also remotely trigger the recognition.

Informed by the pilot study, and to provide more consistent and clear information to the user, we decided to modify the generated feedback. Instead of providing the text following the physical orientation (i.e. top to

Table 2. Details of the participants involved during the user Evaluation and their SUS score.

Case	Age	Gender	Impairment	Acuity	WHO *Score*	SUS
P1	63	Male	Birth Blind	No Light	Blindness (cat=6)	62.5
P2	64	Female	R Blastoma	No Light	Blindness (cat=6)	70
P3	37	Male	Birth Blind	Light	Blindness (cat=5)	65
P4	72	Male	R Pigmentosa	R=0 L=Light	Blindness (cat=5)	57.5
P5	32	Male	Stargardts	L/R=2/60	Blindness (cat=4)	82.5
P6	68	Male	Glaucoma	L=0 R= 6/120	Blindness (cat=4)	72.5
P7	61	Male	M Degeneration	6/100	Severe (cat=3)	75
P8	63	Male	Malformation	6/75	Severe (cat=3)	75
P9	31	Female	Stargardts	L/R=6/60	Moderate (cat=2)	77.5

bottom), we re-organized it based on the font size. For example, text with a larger font size, which often provides more distinctive information, was played at the beginning of the recognition and smaller text at the end.

Finally, to improve the image quality of the system (see Figure 8), we integrated a prepossessing image step that detected if an image was blurred before sending it to the cloud. The system captured a total of 3 consecutive images and calculated the Laplacian variance [45] for each of them. If the Laplacian variance of the image is below a predefined threshold, the image is considered as blurry. The system then submits the image with a higher result to the server. Additionally, to ensure good user experience, we lengthened the UX sound of the camera.

Participants: We invited 9 participants (2 female) aged between 31 and 72 (M=54.55 SD=16.31). We selected participants with different eye conditions: 1 Moderate (cat. 2), 2 Severe (cat. 3), 2 Blindness (cat. 4), 2 (Blindness cat. 5), and 2 Blindness (cat. 6) following WHO vision impairment categories[7]. Further details are in table 2. By evaluating the device for a variety of eye conditions, we get a broader understanding of what eye conditions could benefit most from AiSee.

Procedure: The evaluation was divided into 3 different parts and lasted 1.5 hours.

(1) A grocery item and AiSee was given to the participant. The researcher introduced the task and allowed the participant to practice and familiarize themselves with the device and study procedure.
(2) The participant performed the same recognition task with 20 different items. The items were provided in a random orientation and order.
(3) Finally, the participant completed a standardized usability questionnaire (SUS [12]), together with a semi-structured interview to collect qualitative insights on the apparatus.

Informed by the correlation between the accuracy and the object type during the pilot study, we selected a more diverse subset of items. Our selection consisted of 20 items from a supermarket essentials list, which is a list of commonly purchased items. We avoided refrigerated items, such as fresh or frozen food, to ensure that multiple participants could use the same products.

Task: Participants had to perform two tasks per item: (i) describe the item with as much detail as possible, and (ii) provide the weight or volume details of the product. Within a period of ~2 minutes, the participants were allowed to trigger the device recognition engine as many times as wanted (# Shots). Also, to better understand the extra information the device would provide each participant, we collected a baseline for each item before using the device. The baseline involved performing the same task without the help of any assistive device. During the task, regardless of the participants' baseline knowledge of that specific item, they were only allowed to describe

the object with the information the system provided. This could lead to a situation where the baseline of the participant is higher than the information AiSee provides.

Data Gathering: The descriptions of each item were post-processed by a researcher and quantified as binary (1 correct & 0 incorrect). For task(i) to be correct, the description had to be specific and include at least one item attribute. For example "Tomato Sauce", "Peanut Butter", "Cheese Cracker", "White rice", "High Grain Flour", "Creamy Potato" or "Coconut Milk", etc. If the participant reported chocolate and the item was cooking chocolate, the answer was quantified as incorrect. Similarly, task(ii) was quantified as correct only if the participants provided the exact amount (e.g. 450g). All sessions were videotaped or audio recorded. Additionally, the researcher observed and took notes during the evaluation. SUS forms and open questions were done verbally and entered through Google forms. During the evaluation for participant P4, a connection between the motherboard and the electronics failed, which limited the evaluation of P4 to 7 different items instead of 20. Nevertheless, we were able to collect all the SUS and collect qualitative feedback on the device.

6.2.2 Results.

Task Baseline: Figure 9 top, shows the percentage of items each of the users were able to describe (task i) and provide details of the weight and/or volume (task ii), without the device and with the device.

As expected, baseline knowledge is proportional to the vision of the participants. P9, who has a moderate visual impairment, was able to complete both tasks with almost perfect results. She only faced difficulties with one product (Chocolate Milk Shake), as the font was relatively smaller than other products. P7 and P8, who had severe visual impairment, were also capable of accessing and describing the item without significant complications. However, with P7, his specific disease Macular degeneration, which greater affects the central vision, presented increased difficulties while getting the product's details.

Participants with Blindness guessed the category of some items correctly, such as "Potato Chips" or "Rice". However, they could not provide further details, such as the flavor or type of item. Moreover, on most occasions, the answers were a pure guess, which often results in incorrect responses. For example, P2 and P3 described the item, "Tomato Sauce", as "Shower Shampoo".

Task with the Device: Figure 9 top-orange shows the percentage of items the device was able to identify correctly. These results are independent of the baseline of the participants. Even if the user knows certain information about the item through the baseline, the only information considered is the one the device provides. Figure 9 bottom, represents the number of shots used to extract the information. This number can also be interpreted as the needed "effort" to get the information.

On average, across the 20 grocery items, for task (task i), the device succeeded in 93% of the items, with an average of M = 1.60; SD = 1.09 shots per item. With (task ii), the device succeeded in 67% of the items, with an average of M = 2.17; SD = 1.26 shots per item. These results align with the baseline difficulty between tasks. It also indicates that even though participants exerted higher effort (# shots), this does not correlate with better results.

Usability: Across all participants *(see Table 2)* the SUS score was M = 70.83; SD = 7.9, which is above the 68 points, meaning the system is usable. P4 rated the device's usability consistently lower than other participants, which is likely due to the device breakdown during P4's study session. Moreover, we observed a correlation between the SUS ratings and the eye condition. The average SUS score among totally blind participants (P1 - P4) was M = 63.75 (SD = 5.2). However, for participants with some eye sight (P5-P9) the average score was M = 76.5 (SD = 3.79).

Across participants, SUS question 6 (*"I thought there was too much inconsistency in this system"*), and had the lowest score M = 1.77; (SD = 0.91) – before multiplying by 2.5. Follow-up questions related this inconsistency with the information provided by the user, which changed based on how the image was framed.

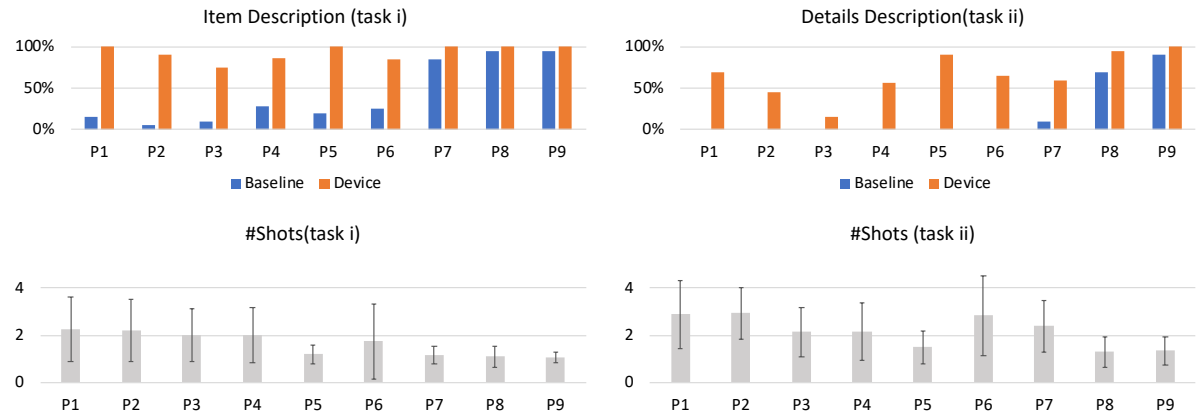

Fig. 9. Top) Percentage of grocery items described correctly, without the device (baseline) and with the device, for each of the tasks. Bottom) Average number of shots performed with the device for each task to extract the information.

Framing: By "Framing," we refer to how the participant captured the information within the camera's field of view. The participants used a number of different techniques to frame the label of the product to the camera. P5, P6, P7, P8, and P9 used their residue eyesight to locate the items' front and back and isolate specific information if needed. With P5 and P6 (both cat. 4), they used their reduced eyesight to locate the packaging information and then triggered the device to receive the device's readout.

The other participants had a stronger dependency on their sense of touch to locate the labels. Furthermore, participants' knowledge about grocery items and text were relevant for the task. P1 had extensive practice in using mobile phone assistive applications, such as KNFB Reader [19] and Seeing-Ai [43], which provided him with practice in finding the correct labels and frame them in front of the camera. P3, who was blind since birth, did not give relevance to the orientation of the item and how the camera framed the item. Also, participants developed this trial-and-error technique to increase performance (e.g., turn the item 90 degrees) and take a shot.

To improve the framing, participants suggested multiple approaches. P6 suggested integrating a laser pointer to the headset so that he could visualize where the center of the camera is located. P1 recommended incorporating a continuous image capturing approach, where the user will keep turning the item, while the device captures multiple images and reconstructs the result. P3 suggested using two cameras, one on each side of the headphones to capture 3D images instead of 2D images.

P5 compared AiSee with existing mobile phone applications and suggested that AiSee will be much more suitable for his daily life, given the nature of his work (outdoors construction). He valued the hands-free aspect of the device. Similarly, P2 compared the system with Seeing-Ai and suggested that for her, our proposed solution made it easier to frame the item. P7 was also impressed by the accuracy of the system, as other applications he used had lower accuracy.

Eye Condition Benefit: As seen in figure 9, participants with better eyesight required fewer shots to extract the information. However, they gained less information from using the device. For instance, P9 (Moderate visual impairment) was able to describe 95% of the items correctly (task i) and provide 90% of the volume/weight (task ii) without the device (baseline). Both of them were elevated to 100% when using the device. Although not quantified, the device also provided the participant with extra information. For example *"No added colour, artificial flavors, and preservatives"*, *"ideal for pickling"* or *"oven bake"*. Similarly, in the case of P8, the device helped him realize

some errors he made while reading the weight using his eyesight. Participants with lower vision (P1, P2, P6) had to perform more shots to extract the information.

Based on our observations, the "benefit" obtained by using AiSee is dictated by the amount of unknown information the device provides. Accordingly, with the current system configuration and task, participants with Moderate visual impairment do not benefit as much as other participants with lower eyesight. Concurrently, the number of shots performed by the user can be used as a metric to understand the amount of effort needed to extract the information.

Taking into consideration these facts, we observed that participants belonging to cat. 4, are the ones that could most benefit from AiSee. For example with P5, whose combined baseline for both tasks was M = 10%. While using the device, it increased to M = 95% with an average of M = 1.35; SD = .57 shots. More details can be found in Figure 9.

Feedback: All participants were able to perceive the audio feedback through bone-conduction without any difficulties, including one of the participants who had hearing loss. At the beginning of the session, participants were asked to adjust the volume and feedback speed.

The system voice was slightly "broken" in some instances, namely after intensive use for ~3 hours (2 participants on a row) without breaks. We associate this issue to an increase in the system's temperature, which triggers a CPU underclocking routine, thus affecting the performance of the Text-to-Speech engine.

A common mistake between some items was the confusion of letter 'g' (grams) and the number 'nine', as well as the letter 'l'(litre) with the number 'one'. This error caused the device to say *"four thousand five hundred nine"* instead of *"four hundred fifty grams"*. Interestingly, participants rapidly spotted this error in the feedback and reported the weight or volume correctly. When later asked, they reported experiencing these errors while using OCR systems.

6.3 Pilot Deployment in the Wild

Next, we wanted to see how users would experience our assistive device, how they would interact with it in an actual supermarket, and reveal the potential impact of our system.

We invited one PVI, who was totally blind (cat. 6), aged 42. The participant did not have any previous experience with the prototype. Before we accompanied him through a local supermarket (*see Figure 10*), we provided a five to ten minute introduction on the project and device. The participant was asked to try the device on his own, picking products of his choice (see Figure 10). To improve the user experience, we also selected some items we assumed the system would identify on its first attempt. Based on our user evaluation, our average success rate for blind participants is ~ 88% (see Figure 9). After the supermarket session, the participant had the chance to discuss his experience, provide suggestions, and articulate wishes. The entire session lasted around an hour.

The participant was capable of interacting with the prototype independently. At least ten objects were taken off the shelves and successfully identified. The participant was amazed by the ability of our proposed technology to assist with shopping. The participant stated: *"I think it would be wonderful for a person with visual impairments to be able to have access to that sort of information"*. The participant also identified the benefit and long-term impact of this technology to increase the PVI's level of independence. He said: *"I am looking forward to the day where I can use the device to increase my level of independence, whether it is shopping or getting information most [things] people would encounter on a day-to-day basis."*

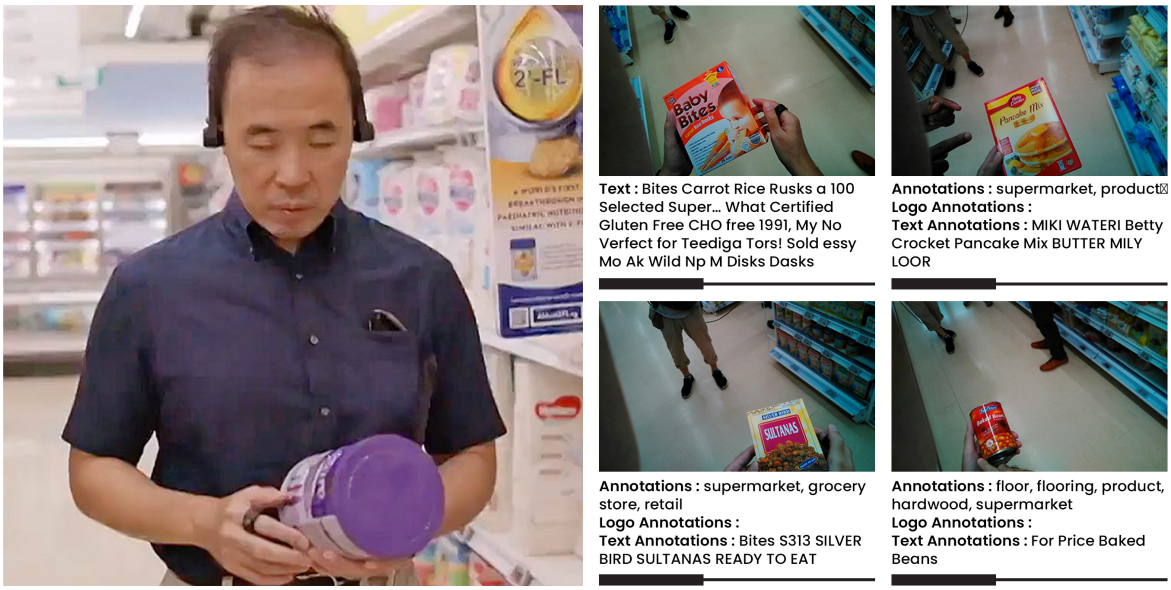

Fig. 10. We conducted an in-the-wild pilot study with one participant using our system in an actual supermarket. On the right side of the image, there are some examples of the products the participant successfully identified.

7 DISCUSSION

7.1 User Evaluation

Eye condition: Different users present different needs. Similarly, the existing eye condition and the user's knowledge are key factors in the usage of assistive technology. In the case of AiSee, we observed a relation between these parameters and the device's usability and performance. For example, participants who were born blind had more difficulties locating the items' labels than participants who lost their sight over time. Interestingly, although those born blind knew where the top and bottom of the item were, they did not turn the item to ensure the text was in the same direction (as done by people not born blind).

Similarly, although users with good residual sight may be unable to read the text present in the label, they can identify the location of this and extract read it by using the device.

Based on these observations, we argue that AiSee will be more useful for users who have lost their sight, or as suggested by P7, for those interested in developing a relationship with technology before losing their existing sight.

Audio quality: Using bone conduction transducers allows the user to listen to information the device provides, while still having access to external sounds, such as conversations or traffic sounds. Participants did not experience difficulties with hearing audio feedback during all three evaluations, despite one participant having hearing loss.

During the evaluation, the researcher could adjust the volume and the speed of the feedback based on the participant's preferences. Despite this, we foresee that if the user is in a noisy environment, such as the metro, they could face difficulties hearing the provided information. The user would need to reduce the external noise by covering their ears or concentrate on the audio in such environments.

Interaction time: After the evaluation, we asked participants to rate whether the system was "too slow" in providing feedback using a 5-point Likert scale. The results (M=3; SD=1.5) indicated that the usability of the system could further improve by providing a quicker reply.

While the device can currently run some basic AI models using Tensorflow lite, the performance from the cloud is considerably better and faster when running larger models. In the past few years, new ASICs have been developed to enable edge computing and run neural networks on devices. Some examples include NVidia Jetson Nano[4] or the Google Coral Dev Board[5]. While such hardware is powerful, well documented, and easily accessible as development boards, it still presents major difficulties when integrating them into wearable devices due to size, power consumption, and market availability. Substantial efforts have been made to enable rapid prototyping with these ASICs on wearable devices. MAIX[6], or Coral M.2, are some examples. We did not use ASIC when developing AiSee, as it was not commercially available at the time. However, we envision future versions of AiSee to embed some of these components to improve the performance and enable on-board processing.

7.2 Image Quality

As reported by previous research [8, 9], image quality is a crucial aspect to perform an image recognition task. Figure 8 summarises the different challenges the device faced regarding the image quality.

Image Sharpness: All three evaluated scenarios were indoor scenarios with artificial light and slightly poor lighting conditions to simulate realistic scenarios (e.g., in a small liquor store). In most cases, the used sensor (Sony IMX135) was able to extract sharp and high-quality images.

Still, a small subset of captured images were blurry (see Figure 8 Blurred - Left). Blurriness is generally caused by a combination of low light conditions (long camera exposure duration) and camera movement (user moving the head). By capturing multiple images and calculating the Laplacian variance [45], we avoided sending blurred images to the cloud. However, we envision the usage of low-light enhancement algorithm [24, 40, 62] for future iterations to improve image quality, while in low lighting conditions. We also encountered that some products generated a light reflection due to their material. These reflections had a considerably negative impact on the recognition algorithm (see Figure 8 Blurred - Right).

Occlusion: As observed in figure 8, the camera captures the ROI of the user without occlusions. However, in some instances, the user's face is on the field of view (see Figure 8 Occlusion). While this does not impair the camera's view, on occasions, it could affect the aperture or exposure of it, which could result in lower quality image and so lower the recognition performance.

During the evaluation, three participants reported wearing glasses (2 while conducting the evaluation). (P2) wore dark glasses to cover her eyes, (P5) wore sunglasses outdoors to protect from sunlight, and (P9) to correct astigmatism. The usage of the device did not interfere with the glasses; however, we did observe a slight occlusion in the camera view among these participants.While in some instances, the user's hair slightly covered the camera's view, the recognition remained unaffected. The camera did not adjust to the focal lens on it.

The placement of the camera was dictated by the relation between discreetness and the field of view of the camera. A better-located camera will potentially take better images; however, it might be more visible to others, creating undesirable social barriers. Further iterations should explore reducing the occlusion by placing the camera slightly further from the users' face while hiding it through a more elaborated form factor.

[4] https://developer.nvidia.com/embedded/develop/hardware
[5] https://coral.ai/
[6] https://www.sipeed.com/

Camera Framing: The framing of the camera has a considerable effect on the performance of the system. AiSee presents two main points of contact of pressure on the head, which are right in front of the ear, where the transducers are placed. However, having 2 points can lead to a rotation on the device. Across sessions, we observed some instances where the apparatus slightly twisted on the user's head, which produced changes to the camera orientation. The camera would point towards the ceiling and capture small parts of the grocery item (see Figure 8 Wrong Framing - Left). Future iterations should introduce a 3rd point of contact to avoid this vertical rotation and increase stability. This contact point could be placed on the back of the neck, together with the electronics to ensure that the device is held consistently at the right angle.

Another difficulty some users faced while capturing the image was the head's alignment with the target object. We designed the device to directly capture the item in front of the user's face without the need to perform any non-natural movement, such as bringing the item close to the head or twisting the head. However, we observed that users understood the camera's location. In some instances, they attempted to compensate the camera by turning the head in a non-natural way in an attempt to improve the camera's framing. However, in most cases, it resulted in poor image quality *(see Figure 8 Wrong Framing - Right).*

ROI - Region of Interest: Extracting the region the user is attempting to recognize (Region of Interest) is essential towards providing a relevant and detailed response to the user. During the iterative process, we defined a natural interaction of pointing to the bottom left of the object in order to extract the ROI. The chosen approach had an important role in increasing the algorithm performance across all the images *(see Figure 8).* Nevertheless, we did observe some instances where the cropping of the image was not successful, and instead harmful to the recognition *(see Figure 8 Cropping - Right).* A possible explanation is that users did not perform the described interaction and pointed with other methodologies. Alternatively, the trained SSD model *(see section* 5.2*)* may have given a false positive *(see Figure 8 Cropping - Left)*, making the system select the wrong ROI and thus extract wrong features of the captured image.

Image Processing: The performance of the system was highly dependent on the type of product targeted. Items with flat surfaces, such as boxes had a significantly higher success rate than other round shape items, such as cans. Furthermore, some products presented a non-standard and complicated text-font, which made the OCR algorithm extract the wrong information. For instance, the algorithm reported "pia" instead of "Pizza", "9" instead of "g", "eleven" instead of "1L", etc. These mistakes are due to the manufacturer's use of a specific font style *(see Figure 8 – Difficult Products).*

We also identified different scenarios where the algorithm provided misleading information to the participants *(see Figure 8 – Confusing Response).* During the pilot study, a can called "Wild Elephant," which has a similar look to a coca-cola can, was wrongly identified a few times and described as coca-cola. These false-positive could be critical in a real scenario, such as buying an unwanted product. Misleading the user can result in severe problems, for example, when the user is allergic to certain ingredients.

7.3 Potential Impact

Many well-performing assistive technologies already exist in the market. However, based on the insights extracted from multiple interviews, we observed a low frequency of use. Often it seems easier to ask others for help than using an assistive tech. We believe that three of the critical points are: (1) Extended functionalities. Users do not face difficulties all the time. It is unlikely users will wear a device all day to complete a task that happens at a specific moment during the day. Therefore, we believe that wearable assistive technology should be presented as an added functionality to already used devices (e.g. headphones or smartwatch) rather than an independent device.

(2) Accuracy on the task, the user wants to know specific information, not general information, (3) Discreetness, users do not want to be targeted as disabled, especially those who are losing their eyesight over time. For example, P7 mentioned that he does not feel comfortable using his smartphone to perform recognition tasks while he is at a shop. However, the user suggested that he could use AiSee at the shop without others noticing.

With AiSee, we target these three aspects. We use the latest artificial intelligence algorithms and enable the user to perform accurate recognition tasks. Our system uses a small ring to trigger the device, and a familiar form factor (Bone-Conduction-headphones). This does not disrupt the user while performing other tasks, while enabling a non-intrusive, quick, and discrete micro-interaction. Finally, we integrate these functionalities into a daily used device, headphones, as they are used all day across multiple applications to play music, perform calls, listen to GPS directions, or interact with the mobile phone. We believe this type of technology can be transformative for people with disabilities, empowering them to perform tasks that currently require assistance. However, before making this technology available to the masses, we understand that there are still vital improvements that need to be addressed, such as a more ergonomic industrial design, a faster processing unit, etc.

8 CONCLUSION

This paper introduced AiSee, which is an assistive wearable interface that helps people with visual impairment overcome daily challenges, such as grocery shopping. Through several interviews, we identified some PVI needs and their pain points with identifying grocery items. In an iterative development process, we developed a high fidelity wearable prototype. The prototype consisted of a bone-conduction headset with a camera capable of processing and extracting features such as text, logos, and labels from the captured images. The device, through the usage of a wearable ring, enabled a simple interaction to perform a recognition task. We evaluated our prototype in two laboratory studies and in-the-wild. Based on this, we are able to provide several insights on the usage of our device and identify some of the remaining challenges to overcome. The emergence of AI-technology is increasing the feasibility of these types of assistive devices, which could potentially create a significant impact on the visually impaired community.

REFERENCES

[1] Aira - Connecting you to real people instantly to simplify daily life. https://aira.io/

[2] Ariadne GPS | Mobility and map exploration for all. http://www.ariadnegps.eu/

[3] Be My Eyes - Bringing sight to blind and low-vision people. https://www.bemyeyes.com/

[4] BlindSquare. http://www.blindsquare.com

[5] TapTapSee - Blind and Visually Impaired Assistive Technology - powered by CloudSight.ai Image Recognition API. https://taptapseeapp.com/

[6] Vision API - Image Content Analysis Cloud Vision API Google Cloud. https://cloud.google.com/vision/

[7] ICD-11 - Mortality and Morbidity Statistics.

[8] Jeffrey P. Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. VizWiz::LocateIt - enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops* (2010-06). 65–72. https://doi.org/10.1109/CVPRW.2010.5543821 ISSN: 2160-7516.

[9] Roger Boldu, Alexandru Dancu, Denys J.C. Matthies, Thisum Buddhika, Shamane Siriwardhana, and Suranga Nanayakkara. FingerReader2.0: Designing and Evaluating a Wearable Finger-Worn Camera to Assist People with Visual Impairments while Shopping. In *Proc. of IMWUT'18*. ACM, 94.

[10] Roger Boldu, Alexandru Dancu, Denys J.C. Matthies, Pablo. Cascon, Shanaka Ransiri, and Suranga Nanayakkara. Thumb-In-Motion: Evaluating Thumb to Ring Microgestures for Athletic Activity. In *Proceedings of the Symposium on Spatial User Interaction (SUI '18)*. ACM.

[11] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2117–2126.

[12] John Brooke et al. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[13] Verena R Cimarolli, Kathrin Boerner, Mark Brennan-Ing, Joann P Reinhardt, and Amy Horowitz. Challenges faced by older adults with vision loss: a qualitative study with implications for rehabilitation. *Clinical rehabilitation* 26, 8 (2012), 748–757.

[14] Michael P Cutter and Roberto Manduchi. Towards mobile OCR: How to take a good picture of a document without sight. In *Proceedings of the 2015 ACM Symposium on Document Engineering*. ACM, 75–84.

[15] D. Dakopoulos and N. G. Bourbakis. Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey, Vol. 40. 25–35. https://doi.org/10.1109/TSMCC.2009.2021255

[16] HIMS International | Blaze ET. http://himsintl.com/product/blaze-et/

[17] Eyra. Horus. https://horus.tech.

[18] Umer Farooq and Jonathan Grudin. Human-computer integration. *interactions* 23, 6 (2016), 27–32.

[19] KNFB Reader App features the best OCR. Turn print into speech or Braille instantly. iOS 3.0 now available. | KNFB Reader. https://knfbreader.com/

[20] Leah Findlater, Lee Stearns, Ruofei Du, Uran Oh, David Ross, Rama Chellappa, and Jon Froehlich. Supporting Everyday Activities for Persons with Visual Impairments Through Computer Vision-Augmented Touch. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (2015) *(ASSETS '15)*. ACM, 383–384. https://doi.org/10.1145/2700648.2811381

[21] Grace Sze-en Foo. Grocery Shopping Assistant for the Blind / Visually Impaired. . http://grozi.calit2.net/files/TIESGroZiSu09.pdf.

[22] James J Gibson. Observations on active touch. *Psychological review* 69, 6 (1962), 477.

[23] Google Brain. TensorFlow Release 1.2.1. https://goo.gl/WZqjLs.

[24] X. Guo, Y. Li, and H. Ling. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Transactions on Image Processing* 26, 2 (Feb. 2017), 982–993. https://doi.org/10.1109/TIP.2016.2639450 Conference Name: IEEE Transactions on Image Processing.

[25] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018-06). IEEE, 3608–3617. https://doi.org/10.1109/CVPR.2018.00380

[26] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.

[27] Step Hear. http://www.step-hear.com/

[28] Wilbert Jan Heeringa. *Measuring Dialect Pronunciation Differences using Levenshtein Distance.*

[29] Rabia Jafri, Syed Abid Ali, and Hamid R Arabnia. Computer Vision-based Object Recognition for the Visually Impaired Using Visual Tags. 7.

[30] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. Supporting Blind Photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '11)*. ACM, New York, NY, USA, 203–210. https://doi.org/10.1145/2049536.2049573 event-place: Dundee, Scotland, UK.

[31] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017) *(CHI '17)*. ACM, 5839–5849. https://doi.org/10.1145/3025453.3025899 event-place: Denver, Colorado, USA.

[32] Ryo Kawamura. RectLabel – Labeling images for object detection for MacOS . https://goo.gl/GVqq9H.

[33] Vladimir Kulyukin and Aliasgar Kutiyanawala. From ShopTalk to ShopMobile: vision-based barcode scanning with mobile phones for independent blind grocery shopping. In *Proceedings of the 2010 Rehabilitation Engineering and Assistive Technology Society of North America Conference (RESNA 2010), Las Vegas, NV*, Vol. 703. 1–5.

[34] Nicholas D Lane and Pete Warden. The deep (learning) transformation of mobile and embedded computing. *Computer* 51, 5 (2018), 12–16.

[35] Patrick E Lanigan, Aaron M Paulos, Andrew W Williams, Dan Rossi, and Priya Narasimhan. Trinetra: Assistive Technologies for Grocery Shopping for the Blind.. In *ISWC*. 147–148.

[36] Sooyeon Lee, Chien Wen Yuan, Benjamin V. Hanrahan, Mary Beth Rosson, and John M. Carroll. Reaching Out: Investigating Different Modalities to Help People with Visual Impairments Acquire Items. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '17* (2017). ACM Press, 389–390. https://doi.org/10.1145/3132525.3134817

[37] Sooyeon Lee, Chien Wen Yuan, Benjamin V Hanrahan, Mary Beth Rosson, and John M Carroll. Reaching Out: Investigating Different Modalities to Help People with Visual I mpairments Acquire Items. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 389–390.

[38] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. *CoRR* abs/1512.02325. http://arxiv.org/abs/1512.02325

[39] Jack M Loomis and Susan J Lederman. Tactual perception. *Handbook of perception and human performances* 2 (1986), 2.

[40] Shiping Ma, Hongqiang Ma, Yuelei Xu, Shuai Li, Chao Lv, and Mingming Zhu. A Low-Light Sensor Image Enhancement Algorithm Based on HSI Color Model. *Sensors* 18, 10 (Oct. 2018), 3583. https://doi.org/10.3390/s18103583 Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[41] Aipoly Fully Autonomous Markets. https://www.aipoly.com/

[42] Denys JC Matthies, Bodo Urban, Katrin Wolf, and Albrecht Schmidt. Reflexive Interaction: Extending the concept of Peripheral Interaction. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 266–278.

[43] Microsoft. Seeing-AI. https://www.microsoft.com/en-us/seeing-ai/.
[44] John Nicholson, Vladimir Kulyukin, and Daniel Coster. ShopTalk: independent blind shopping through verbal route directions and barcode scans. *The Open Rehabilitation Journal* 2, 1 (2009), 11–23.
[45] J.L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000* (2000), Vol. 3. IEEE Comput. Soc, 314–317. https://doi.org/10.1109/ICPR.2000.903548
[46] Roy Shilkrot, Jochen Huber, Roger Boldu, Pattie Maes, and Suranga Nanayakkara. FingerReader: A Finger-Worn Assistive Augmentation. In *Assistive Augmentation*, Jochen Huber, Roy Shilkrot, Pattie Maes, and Suranga Nanayakkara (Eds.). Springer, Singapore, 151–175. https://doi.org/10.1007/978-981-10-6404-3_9
[47] Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. FingerReader: a wearable device to explore printed text on the go. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2363–2372.
[48] Joan Sosa-Garcia and Francesca Odone. "Hands On" Visual Recognition for Visually Impaired Users. 10, 3 (2017), 8:1–8:30. https://doi.org/10.1145/3060056
[49] Lee Stearns, Ruofei Du, Uran Oh, Yumeng Wang, Leah Findlater, Rama Chellappa, and Jon E Froehlich. The Design and Preliminary Evaluation of a Finger-Mounted Camera and Feedback System to Enable Reading of Printed Text for the Blind.. In *ECCV Workshops (3)*. 615–631.
[50] Brian Still and Kate Crane. *Fundamentals of user-centered design: A practical approach*. CRC Press.
[51] Sarit Szpiro, Yuhang Zhao, and Shiri Azenkot. Finding a store, searching for a product: a study of daily challenges of low vision people. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 61–72.
[52] Manoj V Thomas et al. iSee: Artificial Intelligence Based Android Application for Visually Impaired People. *Journal of the Gujarat Research Society* 21, 6 (2019), 200–208.
[53] Tzutalin. LabelImg, graphical image annotation tool on Windows and Linux . https://github.com/tzutalin/labelImg.
[54] Wayne Walls. Comparing image tagging services: Google Vision, Microsoft Cognitive Services, Amazon Rekognition and Clarifai. https://goo.gl/TVdzUR.
[55] Mark Weiser. The Computer for the 21 st Century. *Scientific american* 265, 3 (1991), 94–105.
[56] Samuel White, Hanjie Ji, and Jeffrey P. Bigham. EasySnap: Real-time Audio Feedback for Blind Photography. In *Adjunct Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 409–410. https://doi.org/10.1145/1866218.1866244 event-place: New York, New York, USA.
[57] Help People who are Blind or Partially Sighted. https://www.orcam.com/en/
[58] Jacob O Wobbrock and Julie A Kientz. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.
[59] Katrin Wolf, Anja Naumann, Michael Rohs, and Jörg Müller. A taxonomy of microinteractions: Defining microgestures based on ergonomic and scenario-Dependent requirements. In *IFIP conference on human-computer interaction*. Springer, 559–575.
[60] Meng Ee Wong and Stacey S. K. Tan. Teaching the Benefits of Smart Phone Technology to Blind Consumers: Exploring the Potential of the iPhone. *Journal of Visual Impairment & Blindness* 106, 10 (2012), 646–650. https://doi.org/10.1177/0145482X1210601008
[61] C. Yi, Y. Tian, and A. Arditi. Portable Camera-Based Assistive Text and Product Label Reading From Hand-Held Objects for Blind Persons. 19, 3 (2014), 808–817. https://doi.org/10.1109/TMECH.2013.2261083
[62] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang. A New Low-Light Image Enhancement Algorithm Using Camera Response Model. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 3015–3022. https://doi.org/10.1109/ICCVW.2017.356 ISSN: 2473-9944.
[63] Chien Wen Yuan, Benjamin V. Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M. Carroll. I Didn'T Know That You Knew I Knew: Collaborative Shopping Practices Between People with Visual Impairment and People with Vision, Vol. 1. 118:1–118:18. Issue CSCW. https://doi.org/10.1145/3134753
[64] Tina Chien-Wen Yuan, Benjamin V. Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M. Carroll. I Didn't Know that You Knew I Knew: Collaborative Shopping Practices between People with Visual Impairment and People with Vision, Vol. 1. 118–118. https://doi.org/10.1145/3134753
[65] P. A. Zientara, S. Lee, G. H. Smith, R. Brenner, L. Itti, M. B. Rosson, J. M. Carroll, K. M. Irick, and V. Narayanan. Third Eye: A Shopping Assistant for the Visually Impaired, Vol. 50. 16–24. https://doi.org/10.1109/MC.2017.36
[66] VP Zinchenko and BF Lomov. The functions of hand and eye movements in the process of perception. *Problems of Psychology* 1, 2 (1960), 12–25.