FingerReader2.0: Designing and Evaluating a Wearable Finger-Worn Camera to Assist People with Visual Impairments while Shopping

ROGER BOLDU, Augmented Human Lab, The University of Auckland, New Zealand
ALEXANDRU DANCU, Augmented Human Lab, Singapore University of Technology and Design, Singapore
DENYS J.C. MATTHIES, Augmented Human Lab, The University of Auckland, New Zealand
THISUM BUDDHIKA, Augmented Human Lab, The University of Auckland, New Zealand
SHAMANE SIRIWARDHANA, Augmented Human Lab, The University of Auckland, New Zealand
SURANGA NANAYAKKARA, Augmented Human Lab, The University of Auckland, New Zealand

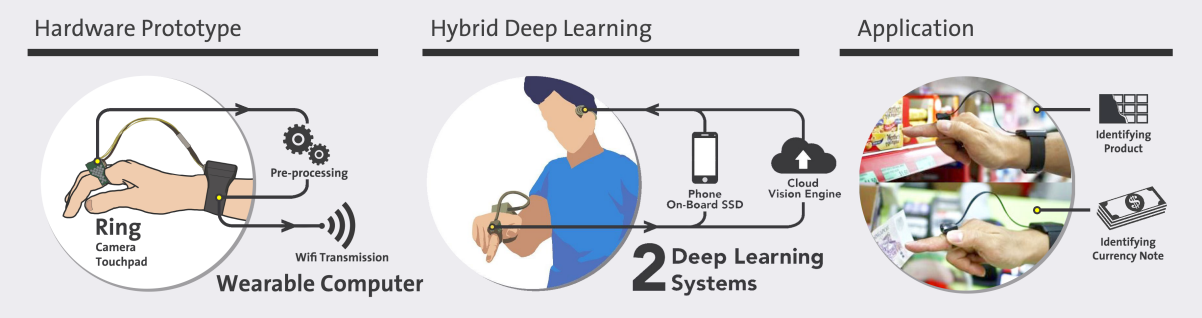

Fig. 1. Our hardware prototype consists of a finger ring, incorporating a camera and a touchpad, as well as a wristband carrying a wearable computer. The image processing is performed either using an on-board deep learning model or using a cloud service, if the object cannot be identified. Our model is trained to understand grocery products, as well as bank notes.

People with Visual Impairments (PVI) experience greater difficulties with daily tasks, such as supermarket shopping. Identifying and purchasing an item proves challenging for PVI. Using a user-centered design process, we understand the difficulties PVI encounter in their daily routines. Consequently, the previous FingerReader model was elevated to a new level. In contrast, FingerReader2.0 incorporates a highly integrated hardware design, as it is standalone, wearable, and not tethered to a computer. Software-wise, the prototype utilizes a deep learning system, relying on a hybrid, an on-board and a cloud-based model. The advanced design significantly extends the range of mobile assistive technology, particularly for shopping purposes. This paper presents the findings from interviews, several iterative studies, and a field study in supermarkets to demonstrate the FingerReader2.0's enhanced capabilities for those with varied levels of visual impairment.

CCS Concepts: • **Human-centered computing** → **Interaction devices**; Graphics input devices;

Additional Key Words and Phrases: Low vision; Wearable technology; Finger-worn camera; Thumb-to-finger Interaction; Accessibility, Hybrid Deep Learning; Assistive Technology; Supermarket Shopping

Authors' addresses: Augmented Human Lab, Auckland Bioengineering Institute, The University of Auckland, 70 Symonds Street, Auckland, 1010, New Zealand. rboldu@ahlab.org, alex@ahlab.org, denys@ahlab.org, thisum@ahlab.org, shamane@ahlab.org, suranga@ahlab.org.

2474-9567/2018/9-ART94 $15.00
https://doi.org/10.1145/3264904

ACM Reference Format:
Roger Boldu, Alexandru Dancu, Denys J.C. Matthies, Thisum Buddhika, Shamane Siriwardhana, and Suranga Nanayakkara. 2018. FingerReader2.0: Designing and Evaluating a Wearable Finger-Worn Camera to Assist People with Visual Impairments while Shopping. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 94 (September 2018), 19 pages. https://doi.org/10.1145/3264904

1 INTRODUCTION

Worldwide 253 million people suffer from low vision [31]. Visual impairment comprises of different conditions and degrees of loss of visual acuity. Depending on the severity of the condition, many of the daily activities of People with Visual Impairments (PVI) are inhibited [3]. Those with visual impairments experience difficulties with specific tasks, such as acquiring products from the supermarket shelf. In these circumstances, PVI depend on others or use tools such as magnifiers or mobile phone applications to assist them in product identification.

One popular phone application to read text by pointing the mobile phone to a document, is KNFB Reader [7]. Recently, deep learning applications are being released such as Seeing AI [28] and Aipoly [27] that are able to recognize items and text in photos taken with the phone. Even though these mobile phone applications are now released, a problem persists: it is difficult to frame a picture without sight [4]. However, due to proprioception skills, PVI can be taught to position a phone camera correctly, as suggested by previous research [4].

An alternative method of using the mobile phone to point to the text and items, would be to employ the gesture of pointing or touching the item that requires reading. Previous research exploring finger-worn text-to-speech systems [38, 41] used computer vision algorithms to perform an Optical Character Recognition (OCR) task and provided the pointed information to the user through audio. Although, this interaction was intuitive and natural, the solution was not wearable and solely focused on OCR task. In FingerReader2.0, we further develop [38] a highly integrated wearable system with on-board Machine Learning algorithms, to explore the task of acquiring a product from a supermarket.

In summary, the contribution of this successor paper is threefold:

(1) Technical development of a standalone, wearable finger-worn prototype, FingerReader2.0, in the form of a small ring with a camera and touch input, which is able to recognize products and notes to help PVI acquire products.
(2) Introduction of the user-centered design process to understand the needs of our user group, as well as discovering challenges and opportunities for the design and evaluation of an assistive finger-worn smart eye.
(3) Compiled insights for designing wearable assistive pointing interfaces for PVI based on interviews, focus groups, and a field study using the FingerReader2.0 prototype inside a supermarket.

2 RELATED WORK

2.1 Shopping Solutions for PVI

A comprehensive study [2] investigating PVI on the difficulties they encountered while taking photographs, found that 28% are related to food or beverage items. Szpiro et. al [43] studied how PVI experience difficulties in locating their desired product on the supermarket shelf. The difficulties they experience 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 [23] is a phone-based system using barcodes and RFID tags to assist PVI in grocery shopping. ShopTalk [29] was a wearable system with a barcode scanner allowing for product search using verbal directions. A smartphone version of these applications were later developed for PVI users [21].

PVI 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. However, using mobile phones to photograph, zoom in, and scan barcodes is problematic. In most circumstances, the resolution is too low and the barcodes are unreadable. Foo [8] developed a grocery shopping assistant that located products and guided PVI to specific locations. She used computer vision and special handgloves for tracking, Wiimote for guidance and 3D audio effects. Lee et. al [24] 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. In contrast, our work utilizes a small wearable ring, a bracelet, and advanced deep learning algorithms for product and bank note identification.

2.2 Low-Vision Aids with Video Output

Zhao et. al [50] created an augmented reality application on a head-mounted display (HMD) that facilitates product search by recognizing the product automatically and utilized visual cues to direct the user's attention to the product. Visual cues were designed according to the visual condition of PVI and evaluated the visual cues with them. Based on their findings, participants preferred using cues to conventional enhancements for product search, as it outperforms best-corrected vision, both in time and accuracy. While this user study focuses on finding the product using visual cues through the head-mounted display, our work focuses on identifying the product using pointing and audio feedback. Zientara et. al [51] created a system composed of smart glasses with an attached camera, that can guide the user towards the desired product through commands (left, right, forward, and back). The system also includes a glove, equipped with a camera that enables *"the viewpoint of what the person is reaching out to hold"*. The solution implements video processing and deep learning on the cloud. The main differences with our work falls within a different scope. We focus on the usability of a ring-camera system that takes single images obtained by tapping on the ring with the thumb. Stearns et. al [40] developed an augmented reality system composed of a finger-worn camera and HoloLens to magnify the text when touched by the finger. However, these approaches are ineffective for those who are fully blind.

2.3 Audio Output for PVI

FingerReader [38] employs a camera worn on the index finger and proposes a novel computer vision algorithm for local-sequential text scanning that allows the user to read single lines, blocks of text or skimming the text with complementary and multimodal feedback. Similarly, HandSight [41] explored the design space of finger-based text scanning. A prototype was developed and evaluated with PVI, exploring how to continuously guide a user's finger across text using three feedback conditions (haptic, audio, and combination of haptic & audio). The two surveys of finger augmentation devices [35, 39] provide an overview of the developments in finger-worn devices, with a classification based on form factor, input, output, action, and domain. The point-and-shoot interaction method is an alternative to line-reading employed in state-of-the-art finger-worn devices, such as the FingerReader [38] and HandSight [41]. Similar point-and-shoot interaction is used by commercial glasses for PVI. OrCam [47] is a small camera, which attaches to glasses. It can read text, describe objects and assist with facial recognition. Horus [6] is a system consisting of a bone conduction headset, which is attached to a small computer and worn on the waist. It uses deep learning to recognize what a user is looking at. KNFB Reader [7], Aipoly [27], and Seeing-AI [28] are some of the most popular text-to-speech and identification smartphone apps PVI use.

2.4 Object Targeting with PVI

The mobile phone applications for PVI share a common problem. The image of the text or document needs to be framed properly, which is a difficult task to perform with low vision [4]. Cutter and Manduchi [4] have compared two feedback modalities on how to orient the phone when taking a picture: guidance with continuous translation

instructions vs. only confirmation when obtaining a good frame. They found that, in addition to phone translation guidance, orientation indications are crucial since 59% of the guidance cases only require simple orientation corrections to have the picture well framed. By using only tactile perception, without vision, a person is able to distinguish two-dimensional and three-dimensional objects by holding and turning them around [26]. Gibson calls this exploratory tactile scanning *"active touch"* [11]. It involves micromotions to explore and measure the object using haptic perception [26, 52]. For PVI in particular, the hands are a core physical interaction channel, as proprioception is well pronounced. Therefore, it seems natural to utilize the high sensing abilities our hands and fingers offer.

2.5 Finger Worn Gesture Interfaces

Thumb-to-finger Interaction: Since the 1990s, the ring's surface was used for cursor manipulation or as a chorded keyboard [9, 34]. Tapping on the surface was also used for replacing the mouse and keyboard [22]; for appliance control [49], as an input device to replace a mouse or keyboard [13, 15, 17, 18, 32], and detecting gestures of a novel vocabulary [20]. More recently, MagicFinger [48] developed tap and gesture input by sensing contact and movement with materials and other fingers.

Finger-worn Touch-pads: Ringteraction [10] proposes a thumb-index interaction, which uses a set of gestures (horizontal and vertical swipes, taps, and rotation) and scenarios (parallel task completion, select+scroll, zoom+pan, focus+context). TouchRing [44] proposes printed electrodes and capacitive sensing on a ring worn on the index finger to enable multi-touch gestures performed with the thumb, palm, and middle finger.

To conclude, our work differs from the related work described above as follows: i) *different application domain:* we focus on mobile usage, with the use case of shopping. ii) *discrete interaction technique:* point-and-shoot interaction is enabled through a touchpad embedded in the ring. iii) *technological advancement:* FingerReader2.0 is standalone, wearable, not tethered to the computer like FingerReader [38] or HandSight [41]. iv) *emergent algorithms:* FingerReader2.0 uses deep learning algorithms to analyze the images and provides information about the item, while previous versions [38], uses computer vision algorithms to perform an OCR task.

3 USER-CENTERED DESIGN PROCESS

We followed a typical system design approach [30] and thus collaborated with our target user group, PVI [36]. The target user group was recruited from the local PVI association. During the process of interacting with PVI, we discovered their daily challenges. Consequently, a device was iteratively developed to assist them in accessing visual information. A total of 26 (5 sighted and 21 visually impaired) different people were interviewed during the user-centered design life cycle.

3.1 Understanding User Needs

To determine PVI's daily challenges and pain points, we conducted a user journey of a typical day with 6 PVI (2 females and 4 males) aged between 21 and 72 years (M=54.1; SD=18.08). The participants were randomly selected from a list of PVI provided by the local association. Their visual impairment conditions consisted of complete blindness (3 participants), retinal pigmentation (2 participants), and diabetic retinopathy (1 participant). Additionally, we also observed 3 visually impaired while attending to an IT class about learning how to use a mobile phone in accessibility mode.

3.1.1 Interviews. The interviews were conducted at the local association for visually handicapped. The interviewer guided the discussion and took notes about the time, location, observation and perspective (thinking, feeling, seeing, hearing, doing), attitude, pain point, and positive aspects while a participant described a typical day. All interviews were video taped.

We extracted the following insights by analyzing the common pain points shared by PVI. All participants, except one (a 21 year old participant) mentioned the use of public transportation (bus or underground train) to get to work, school, or to the association of visually handicapped, to be a major problem in their daily routine. Four participants (diabetic retinopathy, retinal pigmentation and 2 completely blind) mentioned having trouble with grocery shopping. They mentioned that the main problem they experienced is product identification. In order to identify a product, users should be able to use their hands freely to pick up objects and to hold the cane to navigate along isles simultaneously. A wearable device would satisfy this requirement as opposed to a hand-held device, such as a mobile phone, which consistently requires a hand to hold the device.

3.1.2 Observations: Accessibility with Mobile Devices. The PVI attended individual hour-long sessions organized at the local association for visually handicapped. In each session, we trained the participant to use a smartphone and a tablet computer using the built-in accessibility feature. Also, the participants explained their user habit with the current technology they are using. The participants reportedly used Blaze EZ [5] and KNFB reader [7] several times weekly. During these sessions, we observed and took notes about the way PVI learned and used their smartphones (mainly iPhone). One of the key functionalities the user learned, is the *Menu and Hierarchy*. When performing a simple task such as composing an SMS, the user needs to perform several subtasks, such as selecting a contact or composing a message. The resulting hierarchy of tasks and their corresponding options (e.g., a contact is in a list that can be scrolled, inputting text can be done either by scrolling through letters or dictation to Siri) makes it difficult to remember the application state and may lead to confusion. In our opinion, a satisfactory solution to reduce the cognitive load with these hierarchy menus has yet to emerge. Currently, the only option is to return to the main menu and restart the entire process. Reducing the depth of the menu's hierarchy should thus be a priority with future system developments.

3.2 Design Concept

The overarching aim is to enable PVI's to independently access information in a mobile context, as this is a general problem for our target group.

The local community we collaborated with, comprised a majority of people with low vision (71%), which is our target group. Those with low vision are able to partially see objects. They are still able to point towards items that are of interest, in order to obtain information regarding the object.

3.2.1 Defining a Use Case. Based on the users' feedback from the interview, we found that everyday situations in unfamiliar scenarios account for their greatest challenges. In particular, using public transport and buying groceries from the supermarket, was mentioned to be a substantial challenge.

While developments in public transport accessibility for people with visual impairments has occurred [1, 14], little has been done for shopping assistance. However, acquiring food is a basic need and thus enabling PVI to independently purchase products at a supermarket is a major step to (re)gain independence.

From the observational study we learned about technical particularities PVI are struggling with. Therefore, an assistive device should be seamless and not require complex hierarchical menus. Initially, we augmenting the PVI's finger using an *intelligent eye*, which is basically a wearable camera. This enables for the hands to remain free for picking products, holding a basket, or carrying a white cane, while empowering the user with the capability to access visual information independently.

3.2.2 Interaction Concept: Point-and-Shoot. Pointing with the finger to an object is a natural gesture, which we utilize. To prevent constantly flashing the user with verbal information, the user performs a thumb tap on the side of the ring, which takes a picture and analyzes it. We denote this interaction technique as *point-and-shoot interaction*, which implies pointing in the direction of the desired item and triggering the device to get selective information.

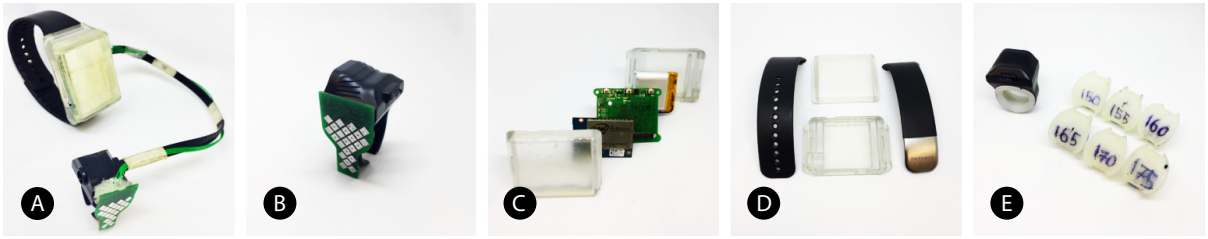

Fig. 2. Prototype components: (A) Wristband connected to the ring (B) Ring with touch-pad (C) Intel Edison, PCB, battery, and wristband 3D-printed casing (D) Casing and straps, and (E) Ring and 3D-printed fittings of different sizes

This microinteraction is different from the interaction method employed in finger-worn devices, such as the previous FingerReader [38] and HandSight [41], in which a tapping gesture to capture the view is not applied. In previous projects, a longer time period is required for the continuous task of reading a line of text sequentially, while the hand and finger must stay in a certain angle, which is difficult for an untrained user.

4 PROTOTYPE

In related work, we can find embedded cameras in several finger-worn devices for line reading such as FingerReader [38], HandSight [41], and gesture input, TouchCam [42]. These devices cannot be worn alone, as they are bulky and connected to a PC. The device's usefulness in a mobile setting is thus limited. However, the prototype's new design overcame these limitations, making it truly applicable for mobile scenarios, such as shopping. We managed to increase the level of integration, added Wifi connectivity, and developed a versatile and scalable platform. For technical reasons, we still require a wristband containing major electronic parts, as well as the battery. This allowed the size of the ring to decrease. *(see Figure 2 - A)*. In contrast to the previous version of FingerReader [38], we now integrated an artificial intelligence onto the finger assisting the user on-the-go in real-time.

Our prototype allows users to simply *point-and-shoot* at products, menus, and a variety of signs to perform a recognition and interpretation task, and hear the results spoken through audio.

4.1 System Architecture

The prototype contains three hardware components:

(1) a ring with an embedded camera and a touch interface (*see Figure 2 - B*). The location of the camera enables the system to capture the image of what the user is pointing at, while simultaneously allowing the user to control the device through the touch interface. The ring is tethered to the second component,

(2) a wristband that contains the processing unit. This processing unit is composed of a system on board, a wireless module (Wifi+BLE), and a battery (*see Figure 2 - C*). The processing unit transmits the captured images to a third component,

(3) a smartphone through Wifi communication. The smartphone performs the image analysis and delivers the information to a user through a Bluetooth Headset or through the phone's speaker.

4.2 Ring Hardware Prototype

Electronics: The ring incorporates a VGA camera module ov7675, with a lens size of 1/9" and 67° aperture. This camera is connected to a main custom made PCB, where there is an SN9C5281BJG DSP from SONIX that controls the CMOS image sensor and transmits the image over a UVC protocol to the external processing unit. On the left side of the ring, there is a 15mm x 20mm touch interface, implemented on an external custom made PCB. This sends the results of the touch input to the external processing unit. The price of the ring electronic components

is as low as 22 USD per unit (producing 200units), making this prototype affordable. Power consumption of the ring is about 4.98mA while streaming 30Fps VGA video image.

Form Factor: The ring is made out of a soft tooling with injected black Acrylonitrile butadiene styrene (ABS) material. Initially, there were two different ring sizes: (i) internal diameter of 19.5mm and (ii) internal diameter of 16mm. Both designs contain a .5 mm gap at the bottom that allows the ring to deform and fit in users up to size 11 (20.6mm). During the interviews, we realized that the two ring sizes were not sufficient to cover the finger variety of the participants. As a solution, fittings adjustment pieces (*see Figure 2 - E*) were modeled and 3D printed with soft material. Starting from the bigger size of the ring (19.5mm), the fittings are inner diameters of the circular ring ranging from 15 to 19mm in .5mm steps.

4.3 Wristband Hardware Prototype

To provide a wearable hands-free experience, the processing unit is physically separated from the small ring with a camera. Hence, the processing unit and battery were moved to the wristband prototype connected with a cable to the ring (*see Figure 2 - A*) prototype. This design allows a wearable hands-free experience, in contrast to holding a smartphone.

Electronics: The wristband processing unit is based on a custom made PCB that operates on an Intel Edison SOM (system on module). This wristband also includes a Dual-core Intel Atom 500MHz processor, 1GB DDR3 RAM, 4GB eMMC flash, Bluetooth 4.0, Wifi, Wi-Fi Direct. The system runs an embedded Linux Yocto 1.1. The wristband is interconnected with the ring through a total of 6 wires. An I2C protocol (SDA and SCL) for the touch interface, D+ and D- for the camera (UVC protocol), power (3V3), and ground. Both the ring and the wristband are powered from a 450mA, 3.7V Lithium polymer battery located in the bracelet. The power consumption tests show that the device can last approximately 3.5 hours.

Software: The wristband serves three main tasks: 1) read from the touch sensor, interpret, and classify gestures. 2) Control the driver of the camera, the wristband configures and captures images from the CMOS camera through the UVC protocol. 3) Communicates to the smartphone through a TCP/IP socket on top of the previously established phone-hotspot. The wristband sends the gestures performed by the user as well as the images captured.

Form factor: The prototype has a similar form factor as Samsung Gear V1. The casing is 3D printed with a multilateral 3D-Printer (Objet 500) and has a size of 35mm x 25mm. The bands used in the prototype are the original parts from the Samsung Gear (*see Figure 2 - D*).

4.4 Phone Application

A hotspot network is opened by the phone to connect with the wristband. Once this communication occurs, an Android phone application receives and analyzes the images from the wristband. Performing the *point-and-shoot* gesture triggers the application to shoot the photo by the finger-worn camera. The received image is analyzed with an on-board identification and an external scene description. Once the image analysis result is returned in plain text, the system synthesizes the audio by using Google text-to-speech engine. The user receives the output through the phone speaker or via Bluetooth headphones.

4.5 Image Understanding and Communication

In order to identify the object the user is pointing at, we evaluated various state-of-the-art image recognition libraries using our hardware prototype. In conclusion, we decided to implement a hybrid approach, which analyzes the captured image in parallel (1) using an on-board deep learning algorithm, as well as (2) sending it to an external cloud vision API. By default, we set priority to our own on-board identification algorithm. When a confident result cannot be achieved, a result from an external cloud service is then requested.

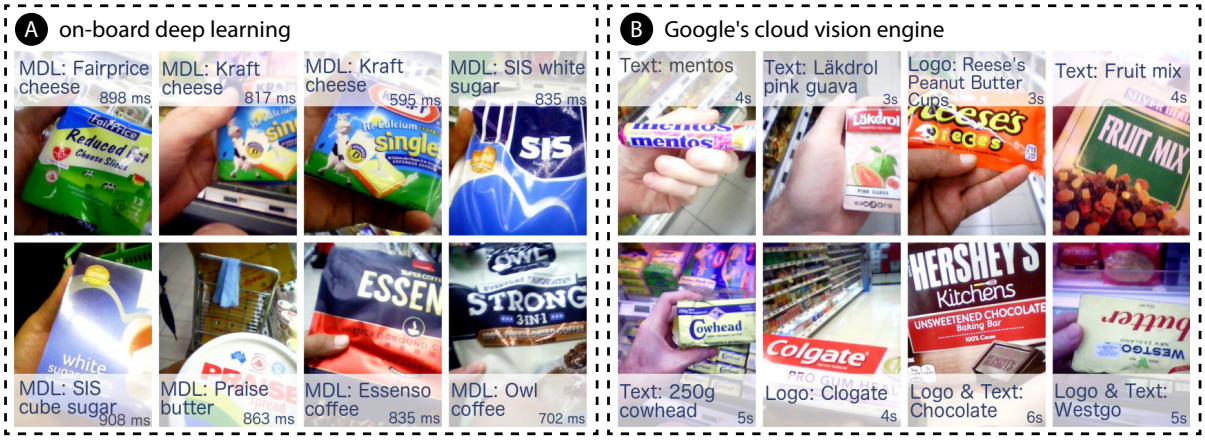

Fig. 3. Successfully recognized images during our field study: A) using our on-board deep learning, yielding low recognition delays <1s, B) using the Google cloud vision engine for untrained products yielded significantly higher delays >3s.

4.5.1 On-Board Deep Learning Object Identification. In order to implement the smartphone on-board identification, we trained our own deep learning algorithm. We explored state-of-the-art convolutional object detectors [16] that can perform object localization and object classification. We used the Single Shot Detector (SSD) [25] which relies on the MobileNet-V1 as a feature extractor. We implemented this architecture by using the TensorFlow 1.21 object detection API [12]. In terms of speed-accuracy trade-off, [16] the SSD architecture results in a fast execution for large objects [16], which outperforms many other architectures on mobile devices. While we had 13 classes in total (4 currency notes and 9 supermarket products), the average processing time for a request is *M*=1.4s; *SD*=.8s (*see Figure 3*).

4.5.2 Labeling and Annotation. For data annotation, the Oxford IIT format was used [33]. It was important that the taken images contain a single object belonging to one of the classes and to have the object straight, to allow it to fit as tight as possible into a rectangular bounding box. The annotation has been implemented using the tools LabelImg [45] and RectLabel [19]. To improve recognition rate, we created a ground truth database, while capturing images in ways that PVI would acquire them: with finger occlusion, bad light conditions, partial product shot, and up side down.

4.5.3 SSD Configuration. The SSD parameters were set as follows: batch-size (24), initial-learning-rate (.0004), use-dropout (true), dropout- keep-probability (.5). During training, we used dropout to avoid the model from overfitting. The learning rate was set to exponential decay and the batch size was set to the original configuration file in Tensorflow (TF). When a new data set was trained, transfer learning was used to initialize our models. We used a pre-trained SSD model from the TF Object detection API (Model Zoo). This model was trained with COCO data set. This reduced training time as it converged quickly, making it possible to use small data sets.

4.5.4 External Scene Description. In case the item was not identified by the on-board identification, the Scene Description mode would analyze the overall image and provide maximum information to help the user acquire a greater understanding of what the user is pointing at. The description of the scene follows these categories in this order: logos, text, general object characteristics (e.g., book, electronics, etc...). The categories and their orders were defined while conducting the interviews with PVI, where we extracted the most relevant categories for the user. Furthermore, feedback on the image framing (*see Figure 4*) was reduced to prevent overloading the user with excess information.

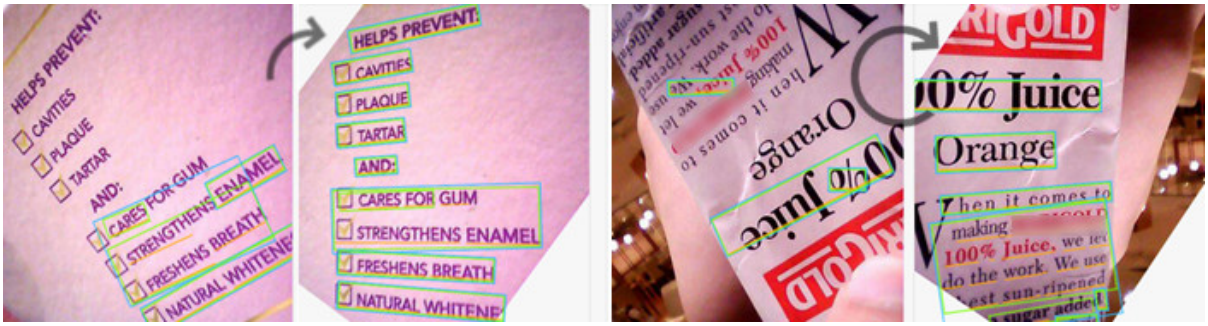

Fig. 4. Original images from the experiment were repeatedly rotated and processed with Google Vision API. The greater number of detected text after rotation is shown by the green bounding boxes.

Based on a previous study [46], we decided to use Google's Vision API. For compatibility reasons, our prototype sends a 640 x 480 JPEG image to Google's cloud vision engine. Although this engine is very powerful, the systems average request time is *M*=3.5s (*SD*=.9s) during our experiments with over 200 requests. Request time includes latencies from sending the image, server processing and receiving the result. While there is a delay, the engine is capable of recognizing a font size of 7 at a 20cm distance, which makes it acceptable. Also, it provides support for more than 50 languages, which can be relevant when identifying ingredients from international products.

4.5.5 User Interface. The audio interface features a synthesized speech implemented by using Google's text-to-speech API. Initially, we implemented a hierarchical audio menu containing different entries: object identification, text reading, and settings. Also, we allowed the user to adjust settings, such as the speed and set a different language. This audio menu was controlled by using a thumb-to-index touch interface. The user was changing the application by swiping up and down, configuring the parameters by swiping right and left, and finally triggering the system by tapping. However, during our iterative design process, we observed that most of the users were uninterested in this feature and used the object identification application only. The main reason for this was the issues some participants experienced when performing gestures. Swiping left and right added an extra layer of complexity to the main task's performance.

Consequently, we developed some adjustments to the user interface: We 1) removed the list of applications, except object identification, 2) simplified the touch input by removing the Logistic Regression classification model and only allowing a single tapping gesture to trigger the recognition, (i.e when the user touches the ring surface, a picture is captured and analyzed); and 3) moved the hierarchical settings menu (voice speed control, language, etc..) to a visual menu on the phone, since its primary use is limited to the initial device setup.

5 ITERATIVE DESIGN PROCESS

5.1 Focus Groups

To gain feedback for the future refinements of the next iteration of the prototype, we organized two focus groups at the local association for visually handicapped. The first group consisted of 4 social workers (sighted), who help PVI with their daily tasks, as well as teach them some of the basic skills, like using the white cane, labelling objects at their home, etc. The second group consisted of 4 employees from the visual impaired institution (3 of them had visual impairments). This second group was composed by experts in accessibility tools for PVIs.

Social workers suggested that PVI would prefer to be independent, such as shop on their own. Some use the Ruby magnifier which is a 4.5" device with a large display. It requires a purse or a bag to carry, which is bothersome. We assumed that a wearable solution like the finger-worn device would increase usability. After the

general questions concerning their daily challenges, we demonstrated the prototype and demoed text reading from a text printed on paper. Additionally, one social worker suggested an attempt to read text from a small chocolate cake package. This suggestion was useful, as it fully captured the FingerReader's capabilities. The speech was played through the speaker of the phone, while it was placed on the table. Their main feedback points can be grouped in the following categories.

Applications: Reading text from receipts and packages, such as ingredients in small font sizes and expiry dates are important. The new generation of ATMs have support systems for PVIs, but require the user to bring their own headphones. It would be useful to know details about the food that the user has in front of them.

Configuration: It would be helpful to add an option to adjust the pace of the speech, as well as to support multiple languages.

Difficulties: On product packages, it is difficult to locate certain information. A bright light, such as a laser, is useful to resolve the difficulties in identifying PVI's placement points. They emphasized that some PVIs can see light and shadows. Some older people experience tremors. They were worried that this would affect the quality of reading the text.

System Design: The ring size should be adjustable to fit any finger size. Connecting it to an earpiece would make it subtle and less embarrassing. However, we were surprised to hear that PVIs did not want to wear earpieces. We believe the main known reason that PVI prefer bone conducted headphones over regular headphones, is that they do not block surrounding sounds. This is crucial, as it directly relates to their safety, particularly when they are walking. We took this into account by using bone conduction headphones in the final experiments.

Feedback: The response time with the detected text is a matter of seconds. It was suggested that knowing the system's continuous state, whether a picture was taken and processed, was an important factor. Furthermore, the users were also uncertain if standing still was a requirement to receive a response.

5.2 Re-entering the User Centered Design Life Cycle

In order to iterate the prototype and improve the user experience, we re-entered the user centered design life cycle. We conducted semi-structured interviews over the span of three months with four PVIs (PI1-PI4). All of them were right-handed, aged between 21 and 66 (M=42.5yrs; SD=22.33yrs), two of them were totally blind, one had Leber's congenital amaurosis, and two of them were female. Every two weeks, we met with one person and incrementally implemented their key suggestions. Two persons were met twice and the sessions were alternated.

The goal was to get initial feedback and quickly improve the usability of the application. The task was to read the text on packages that were arranged on the table, while the user sat down and used the device.

5.2.1 Procedure. We placed the bracelet on the wrist and the ring on the index finger. We introduced the device using one product and provided instructions for its use, namely: i) gestures are performed on the ring touchpad placed on index finger using the thumb; ii) swipe up/down and long press are used in navigation and tap for detection; iii) description of the two states of the system – taking picture and feedback. The user was instructed to hold the object in the left hand while taking pictures with the right hand.

5.2.2 Further Design Iterations.

The interviews were completed based the subject's availability. Not all the participants participated in every iteration. The interviews were conducted at the same association and each interview lasted for a duration between 1 and 1.5 hours.

Iteration 1: The participants appreciated the touchpad menu and that it was intuitive to swipe. However, they encountered difficulties with the long press that was sometimes detected as tap, which was later resolved. The initial menu required two actions to take a picture. Firstly, the user needed to select the read text or object

recognition application. Once the user entered the application, another tap was required to take the picture. This issue was solved by automatically taking a picture after the application was selected. Immediate audio feedback for each gesture was implemented to inform the user that the system has detected input. As the touchpad was placed on the left side of the left side of ring *(see Figure 2 - B)*, the participant raised the concern that it could not be used by left-handed users. The support for scrolling through the feedback character-by-character was implemented. Knowing where the information is located on the package was the biggest challenge. The accuracy of the text recognition was also a problem that was addressed later.

Iteration 2: The task was to identify one product among several similar ones while the user was seated. While performing the task, we noticed that the ring was rather loose. This hindered an accurate camera point and required the middle finger to hold the ring in place. Consequently, for the next session, we created 3D-printed fittings (see Figure 2 - F) that were mounted in the interior of the ring to customize the ring size. Additionally, the user encountered difficulties with pointing correctly towards the text and suggested a similar feedback like KNFB Reader [7]. We noticed an interesting problem that was more pronounced in low lighting. When the thumb touched the ring touchpad, the entire index finger would move. This in turn would move the camera and results in blurry photographs. We addressed this issue by adding an empirically determined delay of .13s prior to capturing the image after the touch input was detected.

Iteration 3: The task was to identify the characteristics of 3 products, with the same packaging, but different flavors or brand. In addition to sitting, we also tested the participants posture when standing. The time taken to identify all juices was 7 minutes. The time decreased to 4 minutes for the second type of packaging. The user suggested to implement a maximum number of images taken on a certain products. If it exceeds 3 tries, another solution or feedback should be presented to the user. The user was stressed and required additional feedback on the quality of the image taken: *"is the text cut or not, is it in the frame?", "Was the fault of OCR?", "I can improve what is under my control."* The user noticed that some geometries of packages have advantages for detecting where the text is located. With boxed objects, it is relatively straightforward to notice that the front includes the name and the flavor, while the back contains the ingredients. However, for cylindrical products such as soda cans, it is difficult to find the flavor and ingredients. We considered the user's feedback and implemented several recommendations, such as reading out detected characters, which form a part of a word, and providing hints of a word continuation and direction.

Iteration 4: The same task from the previous session was selected. We realized the importance of gently tapping on the touch surface when taking a picture, but also pointing it perpendicularly to the text on the object. This way, the image would yield the best recognition results. Another insight was that, the way the package is oriented, makes a difference in the quality of the text recognition results using Google Vision API platform. *Figure 4* shows the differences between original images from the experiment and the results from the rotated images post-experiment. Thus, we decided to add information of the angle of the detected text, which was provided by the API. This information can be used to provide further instruction to rotate the object in the given angle.

5.2.3 Final Improvements. In the last two sessions, we aimed to better understand how PVI pointed their finger to specific parts of the product. Therefore, two types of pointing patterns were evaluated: i) By touching the surface of the object when taking the picture; the bottom of the item can be used as a reference or the edges. ii) Without touching the item when taking the picture; firstly, the middle of the object is identified, secondly, the hand is pulled back 10-20cm, and finally the image is taken. Better results were generally obtained when an overview image was taken by pointing without touching. The main modification after these sessions, was to remove the audio menu and employ a single application that only contains the text reading and object recognition.

6 FIELD STUDY: GROCERY SHOPPING

After a long iterative design process, the technical feasibility was tested and an evaluation of the user's experience with their new assistant technology for grocery shopping was undertaken. (*see also Figure 5*).

6.1 Participants

The recruitment was conducted through the local PVI association, who provided us with randomly selected participants with different visual impairments. Five right-handed PVIs (one female) participated in the study. Two were aged 21 and 22, and three were aged between 57 and 66 (M= 61.3; SD=4.5). They all had experience with the device from previous studies (*see Table 1*).

Table 1. Overview of the participants from the shopping experiment

Case	Age	Visual Impairment	Access Habits	Shopping Habits
P1	61	Short sighted	Magnifier	Weekly
P2	22	Leber's amaurosis	Seeing AI	Never alone
P3	21	Blind Congenital	Seeing AI	Every 3 months
P4	66	Diabetic retinopathy	Voiceover	Monthly
P5	57	Ocular Albinism	Magnifier	Never alone

6.2 Task and Procedure

We met with the visually impaired outside their local supermarket. We first invited them to a cafe, located close to the supermarket, and explained the purpose of the experiment, collected demographics, information regarding their visual impairment, shopping habits, and assistive technologies. The first part of the study was the training phase, in which the participant practiced using the device by detecting currency notes and products (*Figure 5 - A,B*). The second part was to enter the supermarket and allow the participants to freely identify some products independently (*Figure 5 - C, D, E*). However, we also directed them towards specific products we trained with, which we provided during identification (*Figure 5 - F, G*). They were required to choose one product, pay for it at

Fig. 5. Collection of pictures from Supermarket experiment. A: Participant performing money recognition task. B: participant getting familiar with how to use the device. C,D,E: Participant performing the free exploration task. F,G: Participant identifying the pre-trained list of items. H: User paying at the cashier

the counter, where they would confirm the cashier's change (*Figure 5 - H*). The study leader remained with the participants at all times, partially recording certain steps, noting down comments, and also provided assistance if necessary.

6.3 Results

6.3.1 Identifying an item. We analyzed a total number of 236 attempts to recognize 17 items: currency (4 notes scanned twice alternatively) and products (9 types) by the 5 participants. By *attempt* we mean an image that was taken using the finger-worn device by pointing to a particular item and pressing the touchpad in order to recognize it. Each attempt corresponds to a processed image by either the Google Vision API (GV) or our on-board deep learning algorithm (SSD). From the 236 images, 90 were processed by SSD and 146 by GV. SSD had priority over GV, if the image was not recognized by SSD, then the GV result was presented. The participants had, on average, a number of 3.2 attempts per item. From the 236 attempts, 63 (26.6%) were successful, identifying

Table 2. Questionnaire results using a 5-point Likert scale (1=strongly disagree, 5=strongly agree), efficiency of device according to logs, and summary of pointing style

	P1	P2	P3	P4	P5
General					
Helps identify items & read text otg	4	2	4	5	4
Instructions and prompts are helpful	4	4	3	4	4
Speed of the system is fast enough	4	3	4	4	4
Using this system is satisfying	4	2	4	5	4
Using this system is fun	5	3	4	4	4
System Usability Scale					
I would use this system frequently	4	1	4	5	4
System is unnecessarily complex	2	3	2	1	2
The system was easy to use	4	1	5	5	5
Would need support of tech person	2	4	4	1	3
System functions well integrated	2	4	4	1	3
Too much inconsistency in system	2	5	2	1	2
Learn to use very quickly	4	1	5	5	5
System is very cumbersome to use	2	4	2	1	2
I felt very confident using system	4	2	3	5	5
Needed to learn a lot to use system	2	4	1	5	1
SUS result (100)	75	23	73	85	82.5
Efficiency					
Average Attempts per Item	5.3	3	2.7	2.3	2.7
Average Time per Item	12.7s	8.5s	11.6s	10.8s	5.5s
Average Success Rate	0.34	0.54	0.56	0.63	0.48
Pointing Style					
Index Finger Bent	✓	✓	×	✓	✓
Waited for Photo Shot Sound	×	✓	✓	✓	✓
Light Tap	×	✓	×	✓	✓
Pointing at the Middle of the Object	×	✓	✓	✓	✓

the item correctly. For the onboard deep learning method, the number of successful attempts were 57 (63.3%) out of 90, while the rest were false positives (33). From the output of the Google Vision API, only 5 (3.5%) out of 146 requests helped the user to identify correctly the item. To describe further the analysis and present the outcome, we define $SuccessRate = \frac{SuccessfulAttempts}{TotalAttempts}$. The average $SuccessRate_t$ is .51, while the average $SuccessRate_c$ of the correctly identified items is .7. Both are averages across all participants per each of the items the participants recognized. The main reason for a failure in recognizing the items, were the pointing method and external environment conditions.

6.3.2 Paying at the Cashier. Although, the currency was successfully identified by all participants in the training phase. Only one participant did not manage to identify the note when paying at the cashier. The number of images taken until successful identification while paying, ranged between 1 and 3. While the younger participants (P2 P3) required less attempts, the older participants (P1, P4, P5) required substantially more attempts to identify the bank note. On average, younger participants roughly took 2 pictures to successfully identify a note, while elderly participants took an average of 2.4 photos.

6.3.3 System Usability Scale. The average System Usability Scale (SUS) score is 67.5 (SD= 25.7). Table 2 shows the ratings of the participants. SUS has a 5-point Likert scale ranging from 1 (*"I strongly agree"*) to 5 (*"I strongly disagree"*). The highest ratings are for the statements *"I thought that the system was easy to use"* and *"I imagine that most people would learn to use this system very quickly"* (M= 4; SD= 1.73). The lowest rating is for the statement *"I found the system unnecessarily complex"* (M= 2; SD= .71). One participant, a university student suffering from Leber's congenital amaurosis (P2), consistently rated the device lower than all the other participants, thus explaining the high standard deviation. The student's computer science background coupled with a high expectation of the technology explains this result. All other participants rated the device consistently higher.

6.3.4 Video Analysis. Most PVI identified the product type with their hands and later found the brand of the specific product by using the device. Some used haptic perception to feel the shelf's location and products arrangement when freely exploring items. Table 2, bottom summarizes details about how the 5 participants were pointing during the experiment. These details were extracted by analyzing the recorded videos in the supermarket experiment and focused on individual pointing characteristics such as: index finger posture, attention to the notification that the picture was taken, tapping strength, and pointing accuracy. The table 2 shows for P1 a correlation between pointing characteristics and the low performance in terms of success rate and pointing issues.

6.3.5 Suitability for PVI. As indicated in the table 1, all 5 participants of the field study had different visual handicaps. Based on the type of each participant's visual impairment, we observed different techniques to access their surrounding information. For instance, P1, P2 and P4, were able to use their own residual eye sight to identify the location of the information. However, we could not observe any substantial difference in overall usage performance, based on the stage of the users sight. Nevertheless, based on our rather small samples size, we can only make rough estimations on suitability for certain diseases, since these can be pronounced in various levels.

7 INSIGHTS

7.1 Pointing Interaction

Our proposed interaction with a finger-worn camera is based on a natural gesture, namely pointing at a desired item the user intended to identify.

Pointing the sweet spot: There is a combination of factors that lead to the successful identification of an item. This includes, air-pointing, the correct distance, aiming at the item's center, paying attention to the audio

notification, having good light conditions, and a steady hand. These factors will contribute to a quick and successful identification.

Aiming the center: To point at a desired product in a supermarket, PVI have to follow four stages: 1) estimate the shape of the item, which is mainly accomplished by touching it; 2) determine the item's centre, 3) lift the occluding fingers from the item itself, and 4) use the index finger of the opposite hand to point at the center of the item. Although this sounds simple, it can be challenging for PVI. The main problem the video recording revealed was that, the user was unable to aim at the center of the object, resulting in only half of the product being photographed. This problem occurred frequently with P1, who had the highest numbers of attempts and the lowest success rate (*see Table 2*).

Touch-pointing and air-pointing: We distinguish between taking a picture while touching the item and obtaining the image while pointing in mid-air. Users were mostly pointing in mid-air after gaining experience and being confident in determining the right distance to items. P2 mentioned that the main difficulty is *"knowing how to point"* and that *"it takes a few tries to make [it] work"*.

Determine the right distance: The distance to the object being scanned, as well as the font size of the package, are essential factors for a successful recognition. P5 complained that estimating the distance between the object of interest and the camera device is difficult: *"It took some time to learn to use the system efficiently"*. This was not an issue with both of our youngest participants (P2 and P3), who had previous experience with the Seeing-AI app that identifies text, objects, and scenes. Their success rate and the number of required attempts were substantially greater, which suggests that determining the right distance can be learned.

Holding item improves pointing: Our studies revealed that holding an item in one hand and pointing at it with the index finger from of the opposite hand yields a higher control, based on their ability of proprioception. In this manner, the user has a better sense of determining the distance and making a minor adjustment to obtain a good result.

Holding hands still: Performing the tap on the ring's touchpad to trigger image capturing, can result in a blurry image because of the camera's shake. In this case, the user moved the hand after subsequently pressing the touchpad, instead of waiting for the audio confirmation. Furthermore, low light conditions easily increase motion blur. A technical, but computationally expensive solution, would be to continuously capture a video stream and compute a good image based on several samples before and after the tap gesture. However, another effective and simple solution is to instruct the user to press the touchpad rather gently.

7.2 Information Presentation

Point-and-shoot interaction loop: The interaction loop using our *point-and-shoot* technique can be executed quickly and seamlessly, while making use of a natural gesture and providing a responsive audio feedback. A recognition delay of 1.4s for on-board identification does not interrupt the interaction loop. In case the captured image is not properly framed, it requires a retake. Using the vision cloud service yields delays up to 5s, which interrupts the work flow and impacts the user experience. This is also confirmed by a lower SUS score with PVI requiring many attempts.

Limit and summarize information: During our design process, we learned that the information communicated to PVI should be minimal, only containing a brief summary of text, objects, or scene that is recognized by the system. Reading out detailed package description is not preferred by the user, as it is also time consuming. The deep learning architectures; SSD and Google Vision are fast and suitable for this purpose. However, both depend heavily on image framing and a correct pointing.

Individual control preferences: Based on our observational study, users would like to be able to control how the feedback is provided. Some PVI found it important to have control features, such as cancel or skip the recognition of this item or replay the previously recognized item. Offering these functions can reduce frustration and shorten

the work flow. For instance, the system provides a correct image recognition. However, the user failed to recognize the audio feedback initially.

Simple-touch interaction: In contrast, other users were confused by having a great variety of control options (by more complex ring gestures and audio interface). Therefore, we made the design decision to reduce our ring gestures to the simplest interaction possible, a single tap to capture an image. This shifts the workload from menu navigation towards the *point-and-shoot* interaction loop, freeing resources to focus on *pointing to the sweet spot*.

8 LIMITATIONS OF THE CURRENT IMPLEMENTATION

In this chapter we discuss the general technical limitations, as well as application specific limitations based on the evaluation and incremental development of FingerReader2.0.

Requiring haptic perception: Although FingerReader2.0 is designed to identify the product and product details, such as the brand, the list of ingredients, the price, etc., the user is still required to make use of their haptic perception. Usually, the PVI use their sense of touch to locate the item, obtain it's size, shape, and material. This involves micromotions to explore and measure the three-dimensional object using haptic perception [26, 52], also called *"active touch"* by Gibson [11].

Requiring product knowledge: In the next step, the user would assume whether they were holding *tooth paste* or *hand creme*. Assuming this product is new in the store, the FingerReader2.0 would not be able to identify it and therefore would make use of the cloud service, which would read out: *Colgate* (*see Figure 3*). If PVI cannot identify the brand or the shape of the item, our technology is rendered ineffective.

Lighting conditions: The lighting in the supermarket and the reflective material of the product packaging is challenging and can sometimes make the text unreadable because the white light would partially cover the text. Changing the angle of the camera towards the object or the angle of the object could reveal more text.

Object rotation: If the text is oriented correctly, then external computer vision libraries are able to recognize text very reliably (*see Figure 4*). We implemented instructions about correct orientation based on the text angle information provided by the cloud vision engine. However, bent text and experimental fonts remain a challenge.

Consecutive image analysis: If the PVI fail in capturing a good image of the item, an intelligent AI would be able to provide more hints on how to hold the item in a better way. Moreover, using visual computing algorithms could utilize previously shot image fragments in order to form a complete image.

Wide angle lens: Since the location of the camera is comparably close to the object the user is pointing at, making use of a high resulted auto-focus camera is important. Unfortunately, cameras in this micro form factor demonstrate high noise. Although a wide lens would allow more imprecise aiming, we need to consider that a wide lens will introduce additional noise into the algorithms. Also, we would need to train the system with images captured with that specific distortion or equalize the image, which again adds additional noise.

Ergonomics: The users were capable of grasping objects, write phone messages, take money out of the wallet, etc. The only limitation we observed occurred when putting the hand into the pocket. Here, the users faced problems with the height of the ring (*see Figure 2* - B). Since the sensor must be carefully aligned with the finger to avoid occlusions, the ring needs to rest tightly to the finger, making it slightly uncomfortable after an extended usage period.

Training data set: To improve recognition, the photo data set used for training the model, should have also contained images similar to how PVI would capture them, e.g., not well focused, finger occlusion, light reflections etc. While this would substantially improve recognition accuracy, the training phase would have expanded to a longer extent. Here, we faced the trade-off between a convenient training phase yielding lower accuracy and a very time-extensive training phase providing higher accuracy.

Scalability & Emergent algorithms: The existing data set for on-board classification uses around 150 samples for each class, which was manually annotated. This fact limits the scalability of the system, since it requires a

significant amount of human resources and requires a substantial amount of computational processing power. The future goal must be to train more classes with very few samples. Utilizing existing trained models from open source projects, such as Tensorflow Object Detection API [12], would be an option. Furthermore, by using emergent algorithms like meta-learning [37] it is possible to train a model with a smaller amount of images, making the training process faster and scalable. Overall, we believe a cloud-based learning model to be promising. This approach reduces the need of a powerful on-board processing unit, reduces the overall cost, and enables a greater system flexibility.

9 CONCLUSION

This paper introduced the FingerReader2.0, which is an assistive smart eye mounted onto the PVI's index finger. A user-centered design process was utilized, which was useful in understanding the needs and daily challenges of PVI. Grocery shopping was the focus of this research, as it was noted as a major issue for this specific target group. The collaboration with the visually handicapped enabled incremental improvements to the initial prototype. Revised versions of the FingerReader were iterated and finally evaluated with PVI in a field study, the grocery store. While this new field of application is still in development, increasing the number of detectable products by utilizing a general image database of products, must therefore be the core focus of future work.

ACKNOWLEDGMENTS

We would like to thank the *"Singapore Association Of The Visually Handicapped (SAVH)"* and all the participants for their support. The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] BlindWays. http://www.perkins.org/solutions/featured-products/blindways
[2] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. 2013. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2117–2126.
[3] Verena R Cimarolli, Kathrin Boerner, Mark Brennan-Ing, Joann P Reinhardt, and Amy Horowitz. 2012. Challenges faced by older adults with vision loss: a qualitative study with implications for rehabilitation. *Clinical rehabilitation* 26, 8 (2012), 748–757.
[4] Michael P Cutter and Roberto Manduchi. 2015. 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.
[5] HIMS International | Blaze ET. http://himsintl.com/product/blaze-et/
[6] Eyra. 2014. Horus. https://horus.tech.
[7] KNFB Reader App features the best OCR. Turn print into speech or Braille instantly. iOS 3.0 now available. | KNFB Reader. https://knfbreader.com/
[8] Grace Sze-en Foo. 2009. Grocery Shopping Assistant for the Blind / Visually Impaired. . http://grozi.calit2.net/files/TIESGroZiSu09.pdf.
[9] Masaaki Fukumoto and Yasuhito Suenaga. 1994. FingeRing: A Full-time Wearable Interface. In *Conference Companion on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 81–82. https://doi.org/10.1145/259963.260056
[10] Sarthak Ghosh, Hyeong Cheol Kim, Yang Cao, Arne Wessels, Simon T Perrault, and Shengdong Zhao. 2016. Ringteraction: Coordinated Thumb-index Interaction Using a Ring. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2640–2647.
[11] James J Gibson. 1962. Observations on active touch. *Psychological review* 69, 6 (1962), 477.
[12] Google Brain. TensorFlow Release 1.2.1. https://goo.gl/WZqjLs.
[13] Chris Harrison and Scott E. Hudson. 2009. Abracadabra: Wireless, High-precision, and Unpowered Finger Input for Very Small Mobile Devices. In *Proceedings of the 22Nd Annual ACM Symposium on User Interface Software and Technology (UIST '09)*. ACM, New York, NY, USA, 121–124. https://doi.org/10.1145/1622176.1622199
[14] Step Hear. http://www.step-hear.com/
[15] Tatsuya Horie, Tsutomu Terada, Takuya Katayama, and Masahiko Tsukamoto. 2012. A Pointing Method Using Accelerometers for Graphical User Interfaces. In *Proceedings of the 3rd Augmented Human International Conference (AH '12)*. ACM, New York, NY, USA, Article 12, 8 pages. https://doi.org/10.1145/2160125.2160137

[16] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. 2016. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR* abs/1611.10012 (2016). http://arxiv.org/abs/1611.10012

[17] Lei Jing, Zixue Cheng, Yinghui Zhou, Junbo Wang, and Tongjun Huang. 2013. Magic Ring: A Self-contained Gesture Input Device on Finger. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia (MUM '13)*. ACM, New York, NY, USA, Article 39, 4 pages. https://doi.org/10.1145/2541831.2541875

[18] Yuki Kanai, Makoto Oka, and Hirohiko Mori. 2009. Manipulation with Fingers in a 3-D Physical Space. In *Proceedings of the Symposium on Human Interface 2009 on ConferenceUniversal Access in Human-Computer Interaction. Part I: Held As Part of HCI International 2009*. Springer-Verlag, Berlin, Heidelberg, 515–523. https://doi.org/10.1007/978-3-642-02556-3_58

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

[20] Hamed Ketabdar, Peyman Moghadam, and Mehran Roshandel. 2012. Pingu: A New Miniature Wearable Device for Ubiquitous Computing Environments.. In *CISIS*, Leonard Barolli, Fatos Xhafa, Salvatore Vitabile, and Minoru Uehara (Eds.). IEEE Computer Society, 502–506. http://dblp.uni-trier.de/db/conf/cisis/cisis2012.html#KetabdarMR12

[21] Vladimir Kulyukin and Aliasgar Kutiyanawala. 2010. 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.

[22] Alan HF Lam and Wen J Li. 2002. MIDS: GUI and TUI in mid-air using MEMS Sensors. In *Proceedings of the International Conference on Control and Automation*. 1218–1222.

[23] Patrick E Lanigan, Aaron M Paulos, Andrew W Williams, Dan Rossi, and Priya Narasimhan. 2006. Trinetra: Assistive Technologies for Grocery Shopping for the Blind.. In *ISWC*. 147–148.

[24] Sooyeon Lee, Chien Wen Yuan, Benjamin V Hanrahan, Mary Beth Rosson, and John M Carroll. 2017. 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.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2015. SSD: Single Shot MultiBox Detector. *CoRR* abs/1512.02325 (2015). http://arxiv.org/abs/1512.02325

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

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

[28] Microsoft. 2018. Seeing-AI. https://www.microsoft.com/en-us/seeing-ai/.

[29] John Nicholson, Vladimir Kulyukin, and Daniel Coster. 2009. ShopTalk: independent blind shopping through verbal route directions and barcode scans. *The Open Rehabilitation Journal* 2, 1 (2009), 11–23.

[30] DA Norman and SW Draper. 1986. User centred systems design. *Hillsdale, NJ: LEA* (1986).

[31] World Health Organization. Visual impairment and blindness, howpublished = "http://www.who.int/mediacentre/factsheets/fs282/en/", year=2017.

[32] Anala Pandit, Dhairya Dand, Sisil Mehta, Shashank Sabesan, and Ankit Daftery. 2009. A Simple Wearable Hand Gesture Recognition Device Using iMEMS. In *First International Conference of Soft Computing and Pattern Recognition, SoCPaR 2009, Malacca, Malaysia, December 4-7, 2009*. 592–597. https://doi.org/10.1109/SoCPaR.2009.117

[33] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. 2012. Cats and Dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[34] K.R. Prince. 1996. Finger mounted computer input device. https://www.google.com/patents/US5581484 US Patent 5,581,484.

[35] Mikko J Rissanen, Samantha Vu, Owen Noel Newton Fernando, Natalie Pang, and Schubert Foo. 2013. Subtle, Natural and Socially Acceptable Interaction Techniques for RingterfacesâĂŤFinger-Ring Shaped User Interfaces. In *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 52–61.

[36] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[37] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-Learning with Memory-Augmented Neural Networks. In *International Conference on Machine Learning* (2016-06-11). 1842–1850. http://proceedings.mlr.press/v48/santoro16.html

[38] Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. 2015. 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.

[39] Roy Shilkrot, Jochen Huber, Jürgen Steimle, Suranga Nanayakkara, and Pattie Maes. 2015. Digital Digits: A Comprehensive Survey of Finger Augmentation Devices. *ACM Comput. Surv.* 48, 2, Article 30 (Nov. 2015), 29 pages. https://doi.org/10.1145/2828993

[40] Lee Stearns, Victor DeSouza, Jessica Yin, Leah Findlater, and Jon E Froehlich. 2017. Augmented Reality Magnification for Low Vision Users with the Microsoft Hololens and a Finger-Worn Camera. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 361–362.

[41] Lee Stearns, Ruofei Du, Uran Oh, Yumeng Wang, Leah Findlater, Rama Chellappa, and Jon E Froehlich. 2014. 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.

[42] Lee Stearns, Uran Oh, Leah Findlater, and Jon E Froehlich. 2018. TouchCam: Realtime Recognition of Location-Specific On-Body Gestures to Support Users with Visual Impairments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 164.

[43] Sarit Szpiro, Yuhang Zhao, and Shiri Azenkot. 2016. 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.

[44] Hsin-Ruey Tsai, Min-Chieh Hsiu, Jui-Chun Hsiao, Lee-Ting Huang, Mike Chen, and Yi-Ping Hung. 2016. TouchRing: subtle and always-available input using a multi-touch ring. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 891–898.

[45] Tzutalin. 2015. LabelImg, graphical image annotation tool on Windows and Linux . https://github.com/tzutalin/labelImg.

[46] Wayne Walls. 2017. Comparing image tagging services: Google Vision, Microsoft Cognitive Services, Amazon Rekognition and Clarifai. https://goo.gl/TVdzUR.

[47] Help People who are Blind or Partially Sighted. https://www.orcam.com/en/

[48] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. 2012. Magic Finger: Always-available Input Through Finger Instrumentation. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. ACM, New York, NY, USA, 147–156. https://doi.org/10.1145/2380116.2380137

[49] Shengdong Zhao, Pierre Dragicevic, Mark Chignell, Ravin Balakrishnan, and Patrick Baudisch. 2007. Earpod: eyes-free menu selection using touch input and reactive audio feedback. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1395–1404.

[50] Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. 2016. CueSee: exploring visual cues for people with low vision to facilitate a visual search task. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 73–84.

[51] Peter Zientara, Siddharth Advani, Nikhil Shukla, Ikenna Okafor, Kevin Irick, Jack Sampson, Suman Datta, and Vijaykrishnan Narayanan. 2017. A Multitask Grocery Assistance System for the Visually Impaired Smart glasses, gloves, and shopping carts provide auditory and tactile feedback. *IEEE CONSUMER ELECTRONICS MAGAZINE* 6, 1 (2017), 73–81.

[52] VP Zinchenko and BF Lomov. 1960. The functions of hand and eye movements in the process of perception. *Problems of Psychology* 1, 2 (1960), 12–25.

Received May 2018; revised August 2018; accepted September 2018