Human-Computer Synergies in Prosthetic Interactions

John M. Carroll, Sooyeon Lee, Madison Reddie, Jordan Beck,
Mary Beth Rosson

Center for Human-Computer Interaction, Pennsylvania State University
University Park, Pennsylvania 16802 USA
jmcarroll@psu.edu, sul131@psu.edu, mbr5511@psu.edu, beckj@msoe.edu,
mrosson@psu.edu

Abstract. Remote sighted assistance provides prosthetic support to people with visual impairments (PVI) through internet-mediated conversational interactions. In these interactions, PVI broadcast live video to remotely-located, sighted people who engage in speech interactions with PVI to create prosthetic support. These interactions can be quite nuanced, creative, and effective. In this paper, we present a design investigation of remote sighted assistance (RSA) in which computer vision capabilities are integrated into the prosthetic interaction, supporting the human participants in various ways. Our study involved creating design scenarios to identify and concretize future possibilities in order to articulate and analyze design rationale for these scenarios, that is to say, strengths and challenges of RSA integrated with CV. We discuss implications for the design of the next generation of remote sighted assistance.

Keywords: Computer Vision (CV); Design Envisionment Scenarios; Prosthetic Interactions; People with Visual Impairments (PVI); Remote Sighted Assistance (RSA)

1 Introduction

Vision is an important tool for most people to understand the world around them, especially when significant objects are too far away to touch and either do not make sounds, or occur in noisy environments. For 50 years, computer vision technologies have been pursued to augment the experiences of people with visual impairments (PVI). *Computer vision (CV)* seeks to provide direct support to PVI by describing the spaces and things around them. In principle, it can leverage analysis of the PVI's context and current goals. For example, prior work addressed grocery shopping [49], [88], [89], an everyday activity that is constrained and simplified by being carried out in a context of aisles and shelves, and an object taxonomy that is restricted to food and household items. If a computer vision assistive system is primed with a model of shopping activities, it should be able to use the model to better predict and respond to PVI goals and needs.

A complementary type of visual prosthetic is *remote sighted assistance (RSA),* which relies on conversational interactions with a remote person who provides assistance as needed or requested. In these interactions, PVI broadcast a personal

video stream of what is around them, and use talk and text to describe their contexts and needs as remotely-located, sighted humans view the video and talk with the PVI. RSA utilizes pervasive networking infrastructures to create prosthetic conversations. These conversations can be sophisticated and nuanced interactions, even when a PVI and an assistant have not worked together before.

Computer vision prosthetics and remote sighted assistance leverage different kinds of information technology and different human capabilities. We hypothesize that these different technologies and forms of mediated interactions could inspire novel design concepts, leverage one another, and perhaps create new personal and professional opportunities for PVI. In this paper, we address three research questions regarding a possible integrated approach to visual prosthetics: *(1) How could such an integrated approach be better than current video-mediated, sighted assistance services, (2) What could we learn about computer vision-based prosthetics by developing and using an integrated approach, and (3) How could such an integrated approach be better than a pure computer vision approach?*

Regarding question one, it might seem obvious that two sources of visual recognition support would be better than one. However, a brief reflection suggests that two sources of input might make the prosthetic interaction between the remote sighted assistant and the PVI *more complex* and produce redundant analyses. Worse yet, they might produce conflicting and inconsistent results. Thus, we are interested in more specific answers to how an integrated computer-human prosthetic might address current weaknesses in such systems or even produce new desirable experiences. Perhaps the integrated prosthetic would produce synergies in visual support that neither of the two approaches alone would provide. Beyond mere recognition, we are concerned with the *quality of the PVI experience*. Could the integrated approach utilize untapped inherent resources in the context of interaction?

Question two recalls the traditional human-in-the-loop or Wizard of Oz [41] approach to developing recognition technologies. In this approach, a human confederate provides some or all of the recognition capabilities allowing investigation of recognition interactions, and providing guidance to the development of the recognition technology itself. Early Wizard of Oz investigations of speech recognition applications helped to define both the interactions for dictation applications and the requirements for underlying speech recognition technology [59]. Perhaps careful study and analysis of RSA prosthetics could help to articulate new goals and approaches for CV, and along that trajectory identify new possibilities for an integrated approach.

Regarding question three, services that combine computer vision support with live sighted assistance are relatively novel. Pure computer vision systems have strengths, including the speed and accuracy with which they can recognize and describe people, places, and things to users. However, there are also known limitations. Computer vision systems do not articulate context and goals through natural, often highly nuanced, conversational interactions; they do not attempt to empathize with the circumstances and concerns of a person trying to use their output, yet this is one of the most salient characteristics of RSA. Instead, computer vision systems generalize the world; each user would get similar (or the same) descriptions of a place or thing even if that description is meaningful only for some of those people and not others. We are

interested in enumerating the ways in which sighted assistants complement and extend computer vision and vice versa.

This paper is structured as follows. First, we contextualize our work against current trends and insights in assistive technology and computer vision research. Then we present a set of envisionment scenarios that we use as a method for analyzing potential synergies between computer vision and remote sighted assistance. Finally, we discuss these scenarios with regard to our three research questions.

2 Background

Our work leverages three streams of prior research: (1) work in computer vision exploiting top-down identification of meaningful structure in the visual world, sometimes called gist; (2) the current generation of powerful personal devices that provide continuous Internet connection; and (3) better understanding of how PVI routinely engage in conversational interactions with sighted people in carrying out daily activities in the world.

2.1 Computer Vision Approaches

Researchers developing computer vision systems have found the goal of helping PVI to overcome their everyday challenges and difficulties to be extremely motivating [45], [82], [92]. A major focus has been on making visual information more accessible through either object recognition [92] or through processing of tags such as barcodes or RFID [56], [75]. The visual processing of tags or barcodes is an easier problem, simply because of the limited set of visual features that are relevant; nonetheless, important implementation questions remain concerning the devices that are used to gather the visual information, the response time for providing the result to the PVI, and the user experience for the PVI (e.g., how the information is presented and to what extent the process supports follow-up questions or other details).

General-purpose visual object recognition by computer is still a challenging and open research problem, with complexities arising from the nature of the visual feed (e.g., static or video stream); the context surrounding the object of interest (recognizing an isolated object on a shelf is clearly much more straightforward than picking it out of a crowded display); and issues relating to orientation and rotation of arbitrary objects. The relevant algorithms rely on the parsing of the same visual cues that are thought to drive human perception (e.g., shape, color, size). Computer vision scientists working on real world object recognition have been exploring the use of both 3D and 2D models. The 2D models have been in development for a longer period of time, and rely on different frameworks that include GIST [25], [60], SIFT [53], SURF [7], detection of edges and colors [10], [26], [31], and deep learning [43], [47]. The 2D models have been implemented in assistive device research prototypes that help PVI to perform daily tasks such as text reading [38], [69], [72], grocery information reading [82], appliance reading [28], finding an object [21], [67], navigation [30], [34], [71], [76], and face recognition and detection [24].

Along with the extensive body of research on CV for object and text recognition, CV researchers have also been expanding capacities of simultaneous localization and mapping systems (SLAM) and visual odometry systems. To some extent, this has been spurred by the emergence of RGB-D sensors, which has accelerated efforts to provide both indoor and outdoor navigational assistance [29]. The dense depth mesasurement that RGB-D provides has allowed real-time depth sensing and tracking of entities and has enabled real-time 3D localization and spatial map construction. These new functionalities have enabled research [29], [58] on improving dense visual odometry and has led to corresponding improvements in obstacle detection and avoidance during assisted navigation [85], [86].

In addition to research stemming from RGB-D functionalities, algorithmic advances in CV have led to better semantic scene understanding, enabling further research on obstacle detection and avoidance. For example, Stixel World [5] demonstrated the detection of objects' height (from the ground plane) as well as localization in a spatial map. A number of other CV researchers have built on the Stixel World algorithm [62], [66] [51], [78], applying it to real world navigation. The Stixel World semantic segmentation algorithm has also been applied to research on autonomous driving [55], [79], [84].

Combined with RGB-D capabilities, researchers [83], [91] used augmented reality (AR) markers with convolutional neural network (CNN) processing for indoor navigation aids for PVI. Another recent research project [40] developed a future positon prediction model with real-time tracking of pedestrian movement and 3D positioning using RGB-D and CNN and studied pedestrian collsion avoidance for PVI.

With respect to the general task of 3D CV, researchers have been addressing the special challenges that arise in trying to create 3D models of a video stream [32], [39]. In these settings, performance is quickly degraded by motion blur, image resolution, noise, change of light, scale, and orientation; even when solutions to these problems can be found, recognition time continues to be a significant obstacle that keeps the technologies from being useful as practical assistive devices [35]. As a solution to this limitation, some researchers have been investigating a hybrid approach that employs crowdsourcing [17], or that supplements the object recognition process with the processing of RFID tags [56].

An interesting variation of hybrid approaches is the "human in the loop" framework, in which a human is tasked with overseeing and at times intervening or at least validating the work of a computer vision algorithm [18]. Such an approach may mesh well with emerging approaches to computer vision that have been seeking to emulate humans' multi-stage perception process, by first detecting the "gist" of a scene, then focusing in on the visual details that enable object recognition [70]. For example, perhaps a high-level analysis of a 3D visual feed (e.g., indicating where object outlines appear to be, some preliminary guesses about thane object) could be shared and further interpreted by a human trained to "complete" the visual recognition process.

2.2 Ubiquitous Internet and Personal Computing Devices

The birth of Internet and the invention of the World Wide Web (WWW) opened a new age of information access for everyone. The benefit that PVI have from these two inventions in terms of information acquisition seems more significant. The Royal National Institute for the Blind in the U.K. remarked that "the Internet is one of the most significant developments since the invention of Braille… because for the first time ever many blind and partially sighted people have access to the same wealth of information as sighted people and on the same terms," [81]. As broadband has become advanced and prevalent, the Internet has not only provided access to information of many forms but has also played a role in connecting computers to computers, people to people, and things to things. Thus, researchers discuss the evolution of the Internet using three related categories: Internet of Computers, Internet of People, and Internet of Things [22].

Just as the Internet itself has become more ubiquitous, the personal mobile devices used by individuals to access and interact with the Internet (e.g., smartphones, tablets, or other handheld devices) have become more sophisticated. It is now standard that such devices boast a range of sensors (GPS, Gyroscope, accelerometer) and actuators, advanced camera, and powerful processing and storage of data. PVI find smartphones to be particularly useful [80], at least partly because of these devices' easy portability and discreet use patterns. More prevalent usage and the adoption of the smartphone by PVI and the technology advancement of the mobile device together provide a good platform for mobile device-based assistive technology development. As a result, many smartphone assistive applications have become available; these include text reading (talkback, voiceover) [36], object recognition (KNFB reader [42], TapTapSee [74], seeingAI [68], SoundScape [50]), [57], navigation and wayfinding (BlindSquare [14]), [4], [8], [54], [63], [64], [65], [87], obstacle detection [1], and grocery shopping [45], [46], [52], [75].

In addtion to these examples, one particular AT form that has been investigated and improved is RSA-based assistive technology. Internet and mobile device advancement and ubiquity have created a more adquate platform for RSA and conversational prosthetic interactions. This has led to quite a few mobile-based RSA research projects and commercial products. VizWiz projects [11], [16], BeMyEyes [9], and Aira are well known examples. Furthermore, CV technology such as Augmented Reality (AR) and Virtual Reality (VR) have begun to be used in mobile applications to provide people with extended experiences. Incorporating AR/VR into the mobile experience opens up futher possibilities and potential benefits of CV integration into assistive technology for PVI. With the ongoing and rapid advancement of the Internet and the associated opportunities provided by Internet-enabled personal devices, PVI have the opportunity to experience richer information access and sharing than ever before.

2.3 Conversational Interactions

Conversational interactions between a PVI and a sighted assistant involve socializing, clarification, and confirmation. Important characteristics of such conversations include trust, shared understanding of the task at hand, and collaboration. PVI must trust the capabilities and intentions of a sighted assistant, they must agree on what needs to be done and how, and they must both contribute. Continuous communication and clarification can ensure that these characteristics are present throughout an interaction [20]. In addition, the initial establishment of common ground paves the way for successful and respectful collaborations. In a field study of PVI shopping with a helper (family member or friend, or a store employee) we identified three general sources of common ground that can support the conversational interaction [90]. One is *assistance knowledge* (how to assist a PVI, such as paying special attention to navigation routes that are convenient); a second is *interpersonal knowledge* (based on common experience or personal relationship, as when a spouse refrains from asking about product brands because PVI preferences are already known); the third is *domain knowledge* (the practice of shopping, as when a store employee knows immediately which aisle to visit for a particular item). The ongoing assessment and interactions that support common ground are a critical enabler of respectful and smooth conversational support.

In general, assistants could lack key information needed to establish and rely on common ground with the PVI they are assisting. They do not occupy the same place and their vision and hearing is mediated/filtered through a live video feed. They only see what the camera shows. The assistants may not know the PVI's preferences and experience. This creates possibilities of omission (missing information) and misinterpretation. Part of the motivation for the design scenarios we present later is a consideration for how an integrated RSA+CV approach might be able to address some of these challenges [90].

Conversations between sighted assistants and PVI are different from conversations between two sighted people where one is helping the other accomplish a task. Many PVI rely almost exclusively on auditory input to make sense of their surroundings, so they prefer more 'talk' than sighted people would [23]. This difference has been deemed a primary consideration in conversational interaction design. For example, Anam et al. [2] used AI and Google Glass to transmit nonverbal expressions to PVI during face-to-face conversational interactions. In addition, Tanveer et al. [73] investigated the use of a visual-to-auditory Sensory Substitution Device to help PVI understand facial expressions. Their goal is to enhance PVI's social awareness and communication skills.

There are few existing studies of remote sighted assistance as a professional practice due to its novelty. However, our recent interview study of assistants who work for the Aira company (Aira.io) pointed to four broad categories of support: scene description, performance, social interaction, and navigation. At the same time, the interviews highlighted the context dependence of RSA [48]. Notably, we outlined a number of challenges that assistants face in their work. For example, several assistants described trying to navigate PVI down a bustling city sidewalk. Tracking dynamic obstacles and prioritizing information to give to PVI takes effort—effort

that, we believe, could be supplemented with computer vision. We have been leveraging these findings in our ongoing analysis and design envisionment work.

3 Methods

Our design analysis of potential synergies between computer vision and remote sighted assistance involved developing envisionment scenarios that were grounded primarily in our empirical study of the professional practices that are emerging for RSA [48]. Some of the support requirements we learned about in that study are consistent with our earlier analysis of grocery shopping by PVI that pointed to a holistic view of this activity: the process includes conducting an inventory and doing other shopping preparation, traveling to and from the store, completing the needed shopping, and returning home to shelve or otherwise organize the newly purchased items [89]. Thus, it was not surprising to learn that RSA helps PVI to complete household chores and activities. For example, assistants help PVI read recipes and locate ingredients in the kitchen. PVI also work with assistants to navigate public spaces, e.g., finding their way to a bus stop, to a particular retail store, or to an item on a grocery store shelf. Cameras and microphones mediate these interactions. PVI use a smartphone or headset camera to provide visual information and supplement this with verbal information (questions, requests, etc.).

These supportive interactions are more elaborate and nuanced than a simple "*What is that?*" query. Instead, they are conversational, often comprising an ongoing side channel of interaction that extends throughout a complex activity the PVI is engaged in. A particularly interesting example offered by an assistant involved a PVI giving a live presentation for an audience. This requires dynamic coordination and collaboration between the PVI and assistant. In situations like these, the two parties may utilize both verbal and non-verbal cues, including pauses and hand gestures to communicate discreetly. The assistant can help the PVI carry out the presentation by summarizing the points to be made (e.g., as seen on presentation slides), describing images, and mediating interaction with the audience.

For the design analysis presented here, we reconsidered the body of findings from our earlier empirical studies [48], [90] through three lenses: positive design, deficit-driven design, and situated design. Each lens considers a different way to anticipate and analyze how a prosthetic design approach integrating CV and RSA could produce value both for PVI and human assistants.

1. *Positive design* attempts to strengthen previously identified strengths of users, contexts, and existing systems [3]. Computer vision can be employed to emphasize or further develop aspects that assistants and PVI enjoy. For example, one strength of RSA interactions is the virtuous back and forth in which the PVI is confidently and successfully acting and doing things as the assistant is providing advice and guidance, and both are feeling more positive as it proceeds. Envisionment scenarios can deliberately try to probe for new possibilities for such positive designs. In our opportunistic integration of time and space scenario (below) we

suggest the value of greater incidental awareness of opportunistic possibilities for PVI.

2. The second lens we used is *deficit-driven,* or *problem-driven design*, the strategy of mitigating challenges [44]. Interviews revealed tasks and circumstances that make the job of the assistant and the experience of the PVI less smooth [48]. For example, the assistant is missing interpersonal knowledge [90]; he or she does not know the PVI's friends and family, and so cannot "recognize" those faces, even though doing so could enable a more efficient and more natural interaction. If a PVI's user account included a personal library of familiar faces, this challenge might be eased. The challenge could also be mitigated if the CV recognized the familiar faces and reported them to the assistant and/or to the PVI.
3. The third design lens we used was *situated design* [6], based on the argument that we should leverage resources inherent to the context and interaction. Certain useful characteristics already involved can be leveraged more intentionally. For example, interviews of assistants found that PVI often interject pauses in their speech to signal to the remote sighted assistant that they need a situation update or other support [48]. The strategy seems to be particularly useful when the PVI is doing a presentation or leading a conversation as it prevents an interruption of the primary interaction. This is apparently an effective reappropriation of a standard speech convention, but perhaps it could be even further appropriated in organizing an interface that incorporated computer vision. The pauses could be used to trigger a help mode in the dialog.

Typical interactions between assistants and PVI included navigation and wayfinding, object recognition, and text reading. Human assistants described using multiple displays to view video and other information, such as personal details and communication preferences of PVI [48]. They also use text messaging and a Slack channel for interactions with other assistants who are working. In some cases, assistants also access the Internet to search for information, as needed, to support PVI. Informed by these findings, we have developed a set of plausible scenarios to envision how CV might complement RSA in real-world interactions.

4 Scenarios

We elaborate five scenarios depicting ways CV could be used to strengthen remote sighted assistants as they support PVI. These include leveraging CV to: (1) enhance video image quality, (2) recognize faces, (3) navigate unfamiliar spaces and places, (4) track dynamic targets and obstacles, and (5) support opportunistic goal achievement. We frame each scenario in terms of existing research and discuss possible strengths and limitations. Our primary objective is to generate and analyze a diverse set of plausible and concrete design starting points. These could be validated and refined by first sharing them with domain experts, and subsequently developed as interactive prototypes.

4.1 Enhancing Image Quality

Assistants report that it can be tedious to give PVI instructions for positioning the phone camera toward an object with enough detail and clarity that the assistant will be able to provide adequate guidance [48]. Assistants described how much time it can take to achieve an optimal camera position and expressed discomfort at feeling like they were "giving commands" to PVI. They also felt that PVI got frustrated as they tried to respond to the requests for moving the camera, only to receive yet another request for just a bit more adjustment.

Such situations tend to occur more often with assistants and/or PVI new to RSA because neither party has developed the context-dependent communication skills necessary to achieve the desired outcome [48]. For example, an assistant may ask a PVI to move the camera "a little bit to the right," which requires some interpretive work. What does "little bit" mean? How might an assistant modify their directions if a PVI moves the camera too far to the right? This can lead to the repetition of unhelpful, ambiguous directions. This process of adjusting camera position is a necessary step for object recognition and text reading tasks. Bigham et al. [11], [12] and Jayant et al. [37] identified the same issue and tried to address it, but the solutions were limited. We envision the utilization of computer vision technology to address this problem. Consider the following scenario:

> **Scenario 1. Reading Medicine Bottles.** Kelly receives a notification from her phone to take a new, recently prescribed medicine. As she opens the medicine cabinet, she worries about selecting the right bottle and making sure she has the right time and dosage – she currently takes quite a few medications and is not yet confident about this new one. She requests RSA and asks the assistant, Tom, to read the information on the label. So that Tom can get a clear view of the pertinent information, Kelly must position her phone in just the right orientation. But, even with Kelly's effort, Tom can only see a partial address for the pharmacy and no information about time of day or dosage. The text is also tilted because Kelly cannot properly align her phone with the medicine bottle. Instead, Tom uses the built-in CV application to zoom in, grab and align the text so that he can see more of the label to read the relevant information. The application knows to zoom in even further when it detects specialized text (e.g., "mg") so that Tom can see the details easily.

This scenario depicts automatic image enhancement and adjustment of a video feed by computer vision. The proposed functionality solves several problems. First, with the power to manipulate the alignment and clarity of the image, as well as to augment certain aspects of the image (e.g., key pieces of text), Tom and Kelly can circumvent the tedious process of achieving optimal camera positioning. Consequently, Tom and Kelly can complete their task with more efficiency and without causing Kelly to feel like she is failing or being "commanded" to perform a task. These tools would also improve the assistant's experience. A possible downside of this application might be that the image manipulation task distracts the assistant from engaging with a PVI. However, depending on the complexity of the task, an assistant might be able to provide step-by-step feedback to the PVI as they work. For example, Tom could

explain to Kelly what he is doing and give her the pertinent information as it becomes available. Such a design might not be comprehensive or interactive enough to anticipate all questions a PVI might have about what is going on.

4.2 Recognizing Faces

One challenge mentioned by Aira agents is describing the physical appearance of other people. They go out of their way and use a specific vocabulary so as not to use offensive terms and to avoid providing subjective descriptions. Still, describing a person to a PVI on a meaningful level requires considerable attention and time, which limits the amount that the assistant is able to accomplish. Especially in settings where there are many different people, like a dinner with family or friends, this can become impractical. Identifying people can be difficult for PVI in busy or loud places where they cannot clearly distinguish voices, so this is somewhere assistance can be very useful. These considerations led us to the following scenario:

> **Scenario 2. Office Party.** Mary is attending an office party at which she will know some but not most others in attendance. She walks into the party and contacts an assistant. It is noisy, so Mary isn't able to easily identify or locate her friends by voice. She wants to know if any of her friends are around and who else is at the party, but there are lots of people in the room near and far, and it is not practical for the assistant to describe all of them in the level of detail that would make them identifiable. Most of her friends are socializing and did not notice her arrival. The assistant defers to computer vision to analyze the video and recognize faces of people Mary has previously identified to the system as acquaintances. It locates and adds a frame around her friends on the feed – both nearby and farther away – from Mary, and the assistant then tells Mary who is present and where they are.

A facial recognition function could alleviate the difficulty of a sighted assistant sorting through large crowds of people and could allow distant people to be identified. Verbally describing even one person to the point that they are identifiable to a PVI is both time consuming and hard work, especially using the different set of vocabulary required to make descriptions meaningful to someone without vision. With groups that are big and/or loud, computer vision can quickly process visual information that would be much harder for PVI and the remote assistant to gather and coordinate. It also eliminates the need for PVI to rely on other people (who will often be engaged in other activities) to initiate social interaction in such settings; this in turn can enhance the self-inititated interactions and social experiences of PVI.

One potential issue with this feature is the privacy concerns associated with storage of identifiable information about people. Theoretically, faces stored could be recognized anywhere that PVI may use this service, whether it be in public or private and whether the third party wants to be located or not. We would anticipate that this would require advance consent to be negotiated with the acquaintances who will be thus encoded. Additionally, facial recognition software is imperfect, and its capabilities may be limited by movement, camera angles, face angles, lighting, and accessories worn on or around the face, although this presents more opportunities for

computer vision-based interventions. The tradeoffs between the CV assistance and the possible false alarms is an important issue for study.

Facial recognition could be a controversial but undoubtedly helpful contribution of CV to sighted assistance provided to PVI. It would certainly speed up the information PVI receive about people around them by many orders of magnitude and possibly even provide social advantages (e.g., names or other details could also be provided as social cues). It also circumvents the need for an assistant to give a series of detailed descriptions that may not even be relevant, as in many cases, they will not know exactly what (or who) the PVI is looking for.

4.3 Navigating Unfamiliar Spaces and Places

A significant proportion of RSA interactions involve navigation [48]: The PVI needs to move from where he or she currently is to another location. While this kind of interaction has been a success case for RSA support, complications can arise. For example, the PVI needs to describe to the assistant where they are and where they want to end up. This can be challenging if the visual scene lacks distinctive landmarks and/or if the PVI does not know much about the current location, such as when a PVI is visiting a new city or moving around in an unfamiliar place in general.

Assistants called out airports as particularly challenging places to support PVI [48]. Airports present many possible courses of action – many ways to get from one point to another. Sighted individuals are able to assess and make navigation decisions in a number of different ways, such as: move in one direction and re-evaluate, move toward signage to inform further naviation, or ask uniformed staff for help. For PVI, asking for help is the most viable option of these three, but it assumes the PVI can identify an appropriate person for help. The airport challenge led us to envision the following scenario in which the PVI can quickly share location information with a remote assistant, and the assistant can recruit additional digital resources to contextualize this information, for example including detailed schematics of the airport that will help to plan an optimal route.

> **Scenario 3. Airport Check-in.** Sally is at the airport. She is confronted by a vast set of check-in counters extending in both directions, organized by airline. She contacts a remote assistant for assistance, asking for help to find the British Airways check-in counter. The assistant immediately accesses a map of the departure level, and at the same time, locates Sally's phone. British Airways is located leftward in the seventh bank of check-in counters. The assistant guides Sally directly to the check-in counter.

This scenario leverages stored visual guidance information and smartphone geolocation, capabilities analogous to mental models and situational awareness of sighted people. Locating the PVI directly through their mobile device streamlines the initial part of the interaction to help the assistant quickly understand the current location. Accessing the airport map allows the assistant to provide authoritative guidance immediately. A simple refinement of this scenario would pass the PVI's

location information directly to the mapping service, displaying all the location information in a single view for more convenient guidance.

This envisionment strengthens and streamlines the basic navigation scenario that is already known to be a strong point for remote sighted assistance. One possible cost of this envisionment is the integration of maps into the interaction. Assistants may already be too busy and overwhelmed to take on additional information. Another downside is the direct sharing of the PVI's location, though it seems feasible to incorporate the same sort of one-time opt-in dialog structures that have become common in mobile applications or to construe the assistant-PVI relationship as an authenticated/secure interaction.

An upside of this envisionment is that it extends ceteris paribus to many similar navigation scenarios. For example, cities often present indistinct visual information. It is possible to be a block from one's target location and yet not be able to see any indication of that. That kind of problem scenario could occur even more easily for a PVI sharing visual information with an assistant through a relatively limited camera view. However, if the assistant was simultaneously accessing interactive maps and tracking the PVI's location on those maps, then the interactions might proceed more quickly and successfully.

4.4 Tracking Dynamic Targets and Obstacles

PVI using RSA reported that they need to know the distance between themselves and one or more nearby objects, e.g., the distance to a cabinet or interior door, to a small object on a table or counter top, or to a bench or garbage can on a sidewalk [48]. In some cases, it would useful for the PVI to get guidance at the grain of centimeters to avoid knocking over the target object or pushing it away instead of grasping it. A remote sighted assistant can help in these interactions. However, assistants are not always able to provide *enough* distance information because distance is difficult to estimate using two dimensional (video) information.

Particularly challenging distance-to-target issues arise in dynamic environments, for example, in a scenario in which the PVI is moving through a crowded location, such as a grocery store or a public park, in the midst of many other people who are also in motion. The PVI needs to avoid multiple potential collisions with objects that are moving independently through proximal space. The remote sighted assistant can help, but the dialogue could become complicated by the need to refer to multiple other people, each moving along their own distinct trajectory. Providing guidance might involve determining who is the closest threat, and proposing a change in direction to avoid a collision, or proposing that the PVI slow or stop walking altogether to avoid multiple collisions. This led us to envision the following scenario:

> **Scenario 4. Walk in the Park.** Harry is out for a walk in a small city park. It's a nice day, and lots of other people are there, too. Harry requests RSA for directional guidance, explaining that he wants to walk a loop through the park and have the assistant describe the scene to him and help him to adjust his path to avoid obstacles. The assistant sees four people close to and in front of Harry. Two of them are crossing his path but not appearing to be moving closer; one is walking in

a similar direction as Harry; and one is walking toward Harry. The assistant helps him to continuously modify his route and plan ahead but is struggling to determine whether the fourth person is going to collide with Harry, and refers to a CV model that depicts the locations and trajectories of Harry and the people immediately around him. From this analysis, it is clear that the fourth person will indeed cross Harry's path momentarily, and that if Harry turns right or left, he could easily initiate collisions with other people. The assistant immediately suggests that Harry slow down or stop for a moment until the path clears.

This envisionment exploits the capability of CV to construct models of moving objects in three-dimensional space [32]. This capability could augment remote sighted assistance with respect to the challenge of accurately identifying movement trajectories and estimating distances in a dynamic scene from two-dimensional video data. It could also compensate for problems in visual contrast in the video used by the sighted assistant, for example, as a moving object transitions into and out of shade produced by trees. Overall, it allows safer movement in a chaotic, informal social context. This is a case where computer vision addresses a specific challenge for remote sighted assistants and improves the guidance they can provide to PVI.

Beyond supporting collision avoidance, the CV model could be used to enhance understanding and sharing of situational descriptions and rationales. Thereby, it could also help to raise the epistemic level of the prosthetic interaction between the PVI and the assistant.

However, there are possible downsides. For example, in this scenario, the assistant has to perform a more complex role by processing an additional source of information, the model of moving objects, as they guide a PVI. Indeed, because we are positing that the computer vision is more accurate and authoritative in estimating distance-to-target from the video data, this integrated prosthetic design also raises issues of coercion and conflict with automated systems, which are known to undermine human performance [19]. The human assistant would be in the position of regularly overriding their own subjective perceptions in favor of input from computer vision.

In a design envisionment, we cannot be sure these are or would be real problems. However, one reason for envisioning scenarios in the first place is to explore different, *plausible* problems in anticipation of near future engineering and evaluation of RSA prosthetics that incorporate computer vision. In this particular case, the kind of problematic interaction we envision might not occur. We know that humans can be deskilled and demotivated by automation support, but this particular interaction does not exist yet.

4.5 Integrating across Time and Space

During continuous support provision in remote sighted assistance, it is common for an assistant to be "moving around" with PVI as they navigate the physical world. For example, they guide PVI around malls, within retail and grocery stores, and through parks and other public spaces. In these contexts, it is difficult for PVI to activate and pursue opportunistic goals, because they are aware only of what they can sense, what

they may remember from previous activities in that same place, or what the assistant narrates to them. The assistant's main responsibility is to guide the PVI in the agreed-upon task(s). Adding an open-ended, *secondary* task ("tell me anything that seems interesting") may be beyond their capacity.

> **Scenario 5. Dinner Specials.** Royce is walking home from work but wants to run a few errands along the way. He requests RSA and asks the assistant to help him pick up his mail from the box he rents at the downtown post office, make a couple of deposits at his bank, and drop off a prescription at the pharmacy. After visiting the post office and bank, Royce is walking down Main Street toward the CVS to drop off a prescription. Royce hears the hum of a small crowd and asks the assistant about the group of people ahead. The assistant can see a line of people outside the 'Fresh from Home' cafe next to a menu board. The CV application also recognizes a board with text on it and automatically zooms in on the text. The assistant reads out the specials to Royce, who decides he might stop for dinner. First, he wants to compare the menu with the specials at his favorite Marigold cafe, which they passed earlier. The assistant repositions the timeline in the CV model to include Marigold and the CV again zooms in to display its specials. Royce is satisfied; Fresh from Home has more interesting specials tonight.

Most people take for granted their ability to formulate and pursue goals in an opportunistic fashion; it is up to them and the details of their current situation about what and how much "off the track" they are willing to go. But for someone with little or no vision, some of the information that could support this sort of opportunistic behavior may not be accessible at the time and place that it is needed. In the scenario, Royce uses his hearing to recognize the buzz of a crowd, and that sets off an exploratory episode that causes him to eventually change his plans to include dinner. Because some of the information relevant to his new goal was encountered in an earlier time and place, the assistant cannot provide all the information needed. Fortunately the computer vision application maintains some limited visual memory that the assistant can "call up" and investigate on behalf of Royce.

Of course there is no guarantee that "older" visual content will still be available through the computer vision application; the PVI and assistant would then need to decide how important this information is, for example, is it worth revisiting a place (the Marigold) to gather it, or should the new goal be pursued using just what is now available. Nevertheless, the scenario shows how the combination of special perceptual functions (zooming) can combine productively with one possible side effect of processed video content (a sort of ad hoc visual memory).

4.6 Reflection

The design scenarios we identified and analyzed can be viewed through the lens of *universal design* (related concepts include inclusive design, accessible design, and design for all [61]). The key idea of universal design is that the tension between designs appropriate for people with disabilities and designs appropriate for people without disabilities can be less than we might expect. Universal design argues that it

is reasonable to try to address all people, and that designers should remain open to insights on behalf of all people, hence *universal*.

We are impressed that some of our design scenarios appear to be universal in some respects. For example, walking around a busy park with other people running or cycling in various directions, roller skating, or chasing their children can be chaotic and potentially dangerous for a person with visual impairment, but perhaps an elderly or otherwise frail person might feel safer if supported by a remote assistant and potential collision warnings. Similarly, accessing a personal library of faces to identify a person across the room is clearly valuable to a person with visual impairment, but sighted people frequently experience recognition of a face, but some uncertainty about who exactly it is. We are not arguing that there is a perfect and complete confluence between visual prosthetics and worthwhile personal devices, but it appears that some of the capabilities we envisioned could have broader utility.

5 Discussion

We were inspired by expressions of interest by PVI we work with in remote sighted assistance to think about how computer vision and RSA could be combined and mutually leveraged. In this paper, we took a design analysis perspective, envisioning a series of specific interaction scenarios with respect to how an integrated prosthetic could address challenges or realize new opportunities in supporting PVI. These scenarios were grounded in empirical studies of current remote sighted assistance [48].

In human-centered design, a well-established approach to understanding how people might experience and appropriate new technologies, particularly recognition technologies, is the Wizard of Oz [41]. In this paradigm, a user interface is presented to a participant, but "behind the curtain," there is a person who uses their human faculties to provide the recognition capability. Indeed, recent research has employed this approach for prosthetic interactions by having a research team confederate review camera feeds in real time to identify visual content [48]. From this standpoint, RSA takes the Wizard of Oz paradigm even more seriously, in that the sighted human assistant not only does the recognition but also participates in a conversational interaction with the PVI about the content.

5.1 Enhancing RSA with CV

Our investigation considered how integrating RSA with CV might provide a better assistive service than current video-mediated, sighted assistance services. This involved, for example, integrating information from general online information systems, such as airport schematic databases, into RSA conversations. There could also be more personalized databases, such as libraries of friends or co-workers faces. In these examples, information that a sighted assistant might not otherwise have access to becomes available and, thus, potentially strengthens their capacity to provide assistance. In the Airport Check-in scenario, for instance, the assistant can

provide much more fine-grained navigational guidance if they have access to an airport map. In the Office Party scenario, they can identify which co-workers are present in the room. A possible downside to computer vision in these scenarios is that they create more work for the assistant.

Our design scenarios each raise specific claims about how CV might enhance RSA. These envisionments provide concrete guidance for further engagement with PVI and for designing interactive prototypes. They also provide guidance for refining the requirements that direct research in computer vision. They help us more specifically focus attention on how visual prosthetics that integrate RSA and CV provide value to PVI and to human assistants.

5.2 Improving CV Requirements Through RSA

Our second research question asked what we might learn about CV-based prosthetics by developing and using an integrated approach. Developing plausible, near-future RSA scenarios involving CV helps to articulate questions and issues that can guide CV research. For example, our design analysis identified ways to augment RSA interactions in ways that leverage Internet services and information beyond what CV alone could provide. The Airport Check-in scenario assumes that the remote assistant has access to stored airport schematics. Human vision, broadly understood, depends on mental models of places and on visual reasoning and memory. These capabilities are not part of the core visual system, but they are crucial to making the visual system effective in real interactions in the world. The Walk in the Park scenario leverages CV to construct models of dynamic targets and obstacles, which require the assistant to interpret and translate the model into concrete directions to a PVI so that they avoid a collision or reach the desired dynamic target.

Our scenario analysis emphasizes the degree to which effective human vision depends on encyclopedic information, such as knowledge of how spaces are organized in general as well as mental maps of spaces, including very specific spaces like one's home. In our scenarios, vision is dynamic, viewpoints are constantly changing, and various kinds of objects are moving through the shared immediate space. The sighted assistants are much more valuable for their ability to recognize colleagues' faces or precisely which check-in counter a PVI needs to approach. Each of these two topics could serve as a seed for future CV research projects.

A longstanding goal for CV research is the analysis of static displays. CV can recognize, with some accuracy, people, their facial expressions and age, objects and text, landmarks, and so forth. However, effectively tracking and predicting the movement of dynamic targets and obstacles remains a challenge, and this is precisely the type of progress necessary in order to realize a scenario like Walk in the Park. Moreover, Walk in the Park may be typical of many other interactions in which PVI navigate offices, buses, crowded stores and sidewalks, all of which put them into dynamic systems with lots of moving parts (e.g., people, animals, cars, bicycles, etc.) [49]. Assistants report that they are overwhelmed at times by the volume of dynamic information that must be processed to enable them to safely guide PVI to their destinations [48]. Knowing where an object is at any one moment is of little value if it

is traveling through space. There is a need to prioritize CV research that aims to track dynamic objects, modeling and predicting movement trajectories.

It is also important to consider how CV may be integrated with human stakeholders. For example, recognizing the presence of a human face is not enough to determine whether and how that face could be meaningful to PVI. It is important to determine whose face it is, how that person is related to the PVI, and how they are relevant at the moment. Facial recognition per se is insufficient to convey these kinds of answers. In our scenarios, the assistants are much more valuable to PVI if they can recognize what is important and why.

For RSA applications, CV also needs to be usable for the everyday person. There are a few groups that have access to and frequently utilize CV, but it is not common in the general population to have experience with such technology. If someone without technical expertise is to launch and interact with computer vision while providing real-time support to PVI, making adjustments to fit their needs along the way, there needs to be a simple, user-friendly interface through which they can easily manipulate the software while multitasking. Control over the computer would improve the RSA in any of our scenarios. This is also a promising research direction to pursue as CV use expands and becomes more accessible and useful to the everyday person.

5.3 CV-enhanced RSA versus Pure CV

Our third research question asked how an integrated approach might be better than a pure CV approach. Our investigation identified scenarios in which remote assistants interact with PVI to refine or modify goals during an interaction. The actors carry out a conversation through which they achieve vision. For example, the Walk in the Park scenario raises the possibility of a constant stream of dynamic targets and obstacles, which could stress a collision prediction algorithm. In such cases, a remote assistant may need to be prepared to "override" the recommended path or action. On the other hand, the Dinner Specials scenario illustrates the value of combining CV with a sighted human assistant. A sighted assistant may recognize the possibility of an unplanned opportunity like stopping for dinner at a new cafe whereas a pure CV system may not. Such an interaction is not merely one of "recognition," instead it is improvisation and problem solving.

RSA protocols – the kinds of guidance interactions that occur between human assistants and PVI – are constrained by guidance policies and by the strengths and limitations and the information the human assistant receives or can access. As an example of such a policy constraint, Aira agents do not provide guidance during the time a PVI is crossing a street. Guidance is provided immediately before the PVI enters an intersection, and immediately after, but not during the crossing. The reasoning for this policy is that the PVI should concentrate attention on their own mobility skills and not be distracted by external guidance. This sort of constraint on RSA protocols is explicit doctrine. Aira agents learn the street crossing rule as part of their job training [48].

Constraints arising from characteristics of the information available to the human assistant are typically more implicit. As an example, it is difficult to read small print on a medicine bottle, especially when the PVI holds it in at an unusual angle. It might be difficult to diagnose a commotion happening behind the PVI (based on ambient sound) unless the PVI pointed their camera in the direction of the commotion. We hypothesize that our envisionment scenarios can be used to "probe" PVI and experienced human assistants (such as the Aira assistants) for more specific implicit knowledge about RSA protocols [15], [33].

The RSA protocol leverages the design metaphor of walking with a human partner [13], [88], [89], [90]. This makes such a service easy to understand for PVI. It engages pre-existing conversational skills from both the assistants and PVI. Indeed, the human partner metaphor can be seen as a design strategy for "explainable AI" [27]. In current human-to-human implementations of RSA, assistants and PVI do a lot of explaining to one another – explaining goals, constraints, and situations [48]. It is becoming widely recognized that effective AI, including CV, must be accountable and explain itself to humans [77].

As we contemplate integrating computer vision capabilities to enhance the resources available for providing guidance to PVI, this also raises the question of whether the human partner metaphor should be critically rethought. One concrete question is whether computer vision analyses could be used to better inform the assistants or directly inform the PVI: If the computer vision informs the assistant, the assistant would be better informed but also might have a rather different job; instead of primarily monitoring the audio and video feeds from the PVI's device, they might have several other displays to view and/or feeds to monitor. If the computer vision directly informs the PVI, the RSA interaction protocols would be entirely different in that there would be two sources of guidance and potentially multiparty conversations about courses of action.

6 Conclusion

We presented a series of envisionment scenarios to explore ways of integrating computer vision (CV) with video-mediated, remote sighted assistance (RSA) for people with visual impairments. Scenarios support the exploration and development of plausible, near-future design and research problems. These scenarios explore how computer vision can complement sighted human assistants and provide a foundation on which to build a research program to realize these envisionments. We suggested how human assistants can provide better support with regard to: (1) enhancing video image quality, (2) recognizing faces, (3) navigating unfamiliar spaces and places, (4) tracking dynamic targets and obstacles, and (5) supporting opportunistic goal achievement. We acknowledge that, in each scenario, computer vision could overburden assistants who already have a lot of information to process in order to provide support to PVI. These scenarios are a concrete step in understanding how to integrate RSA and CV approaches to visual prosthetics, how to identify new requirements for computer vision through creating and deploying integrated

prosthetics, and how to understand the potential benefits of integrated approaches relative to current CV and RSA initiatives.

Acknowledgments. This work was supported by US National Science Foundation (award 1317560). Jordan Beck is now at the Milwaukee School of Engineering in Milwaukee, Wisconsin, USA.

References

1. Amemiya, T., & Sugiyama, H. Haptic handheld wayfinder with pseudo-attraction force for pedestrians with visual impairments. In Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility (pp. 107-114). ACM (2009)
2. Anam, A. I., Alam, S., & Yeasin, M. Expression: A dyadic conversation aid using google glass for people with visual impairments. In Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication (pp. 211-214). ACM (2014)
3. Avital, M., Boland, R. J., & Lyytinen, K. Introduction to designing information and organizations with a positive lens. *Information and Organization*, *19*(3), 153-161 (2009)
4. Azenkot, S., Ladner, R. E., & Wobbrock, J. O. Smartphone haptic feedback for nonvisual wayfinding. In The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility (pp. 281-282). ACM (2011)
5. Badino, H., Franke, U., & Pfeiffer, D. The stixel world-a compact medium level representation of the 3d-world. In Joint Pattern Recognition Symposium (pp. 51-60). Springer, Berlin, Heidelberg. (2009)
6. Ball, L. J., & Ormerod, T. C. Structured and opportunistic processing in design: A critical discussion. International Journal of Human-Computer Studies, 43(1), 131-151 (1995)
7. Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. Speeded-up robust features (SURF). Computer vision and image understanding, 110(3), 346-359 (2008)
8. Behmer, J., & Knox, S. LocalEyes: accessible GPS and points of interest. In Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (pp. 323-324). ACM (2010)
9. BeMyEyes https://www.bemyeyes.com/
10. Bermudez-Cameo, J., Badias-Herbera, A., Guerrero-Viu, M., Lopez-Nicolas, G., & Guerrero, J. J. RGB-D computer vision techniques for simulated prosthetic vision. In Iberian Conference on Pattern Recognition and Image Analysis (pp. 427-436). Springer, Cham (2017)
11. Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., ... & Yeh, T. VizWiz: nearly real-time answers to visual questions. In Proceedings of the 23nd annual ACM symposium on User interface software and technology (pp. 333-342). (2010)
12. Bigham, J. P., Jayant, C., Miller, A., White, B., & Yeh, T. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (pp. 65-72). IEEE. (2010)
13. Blackwell, A. F. The reification of metaphor as a design tool. ACM Transactions on Computer-Human Interaction (TOCHI), 13(4), 490-530 (2006)
14. BlindSqure. http://www.blindsquare.com/
15. Boehner, K., Vertesi, J., Sengers, P., & Dourish, P. How HCI interprets the probes. In Proceedings of the SIGCHI conference on Human factors in computing systems(pp. 1077-1086). ACM (2007)

16. Brady, E. L., Zhong, Y., Morris, M. R., & Bigham, J. P. Investigating the appropriateness of social network question asking as a resource for blind users. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 1225-1236). (2013)
17. Brady, E., Morris, M. R., Zhong, Y., White, S., & Bigham, J. P. Visual challenges in the everyday lives of blind people. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2117-2126). ACM (2013)
18. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., & Belongie, S. Visual recognition with humans in the loop. In European Conference on Computer Vision (pp. 438-451). Springer, Berlin, Heidelberg (2010)
19. Carr, N. The glass cage: Automation and us. (2013)
20. Carroll, J. M., Neale, D. C., Isenhour, P. L., Rosson, M. B., & McCrickard, D. S. Notification and awareness: synchronizing task-oriented collaborative activity. International Journal of Human-Computer Studies, 58(5), 605-632 (2003)
21. Chincha, R., & Tian, Y. Finding objects for blind people based on SURF features. In Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on (pp. 526-527). IEEE (2011)
22. Coetzee, L., & Olivrin, G. Inclusion through the Internet of Things. In Assistive Technologies. InTech (2012)
23. Coroama, V., & Röthenbacher, F. The Chatty Environment–providing everyday independence to the visually impaired. In Workshop on ubiquitous computing for pervasive healthcare applications at UbiComp (2003)
24. Denis, G., Jouffrais, C., Vergnieux, V., & Mace, M. Human faces detection and localization with simulated prosthetic vision. In CHI'13 Extended Abstracts on Human Factors in Computing Systems (pp. 61-66). ACM (2013)
25. Friedman, A. Framing pictures: The role of knowledge in automatized encoding and memory for gist. Journal of experimental psychology: General, 108(3), 316 (1979)
26. Geusebroek, J. M., Van den Boomgaard, R., Smeulders, A. W. M., & Geerts, H. Color invariance. IEEE Transactions on Pattern analysis and machine intelligence, 23(12), 1338-1350 (2001)
27. Gunning, D., Stefik, M., Jaesik Choi, Miller, T., Stumpf, S. & Yang, G.-Z. XAI—Explainable artificial intelligence. Science Robotics, Vol 4, Issue 37 (2019)
28. Guo, A., Chen, X. A., & Bigham, J. P. Appliancereader: A wearable, crowdsourced, vision-based system to make appliances accessible. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2043-2048). ACM (2015)
29. Gutiérrez-Gómez, D., Mayol-Cuevas, W., & Guerrero, J. J. Inverse depth for accurate photometric and geometric error minimisation in RGB-D dense visual odometry. 2015 IEEE International Conference on Robotics and Automation (ICRA) (pp. 83-89). IEEE. (2015)
30. Horne, L., Alvarez, J., McCarthy, C., Salzmann, M., & Barnes, N. Semantic labeling for prosthetic vision. Computer Vision and Image Understanding, 149, 113-125 (2016)
31. Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J., & Zabih, R. Image indexing using color correlograms. In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on(pp. 762-768). IEEE (1997)
32. Hub, A., Hartter, T., & Ertl, T. Interactive localization and recognition of objects for the blind. In California State University, Northridge Center on Disabilities' 21st Annual International Technology and Persons with Disabilities Conference (2006)
33. Hutchinson, H., Mackay, W., Westerlund, B., Bederson, B. B., Druin, A., Plaisant, C., ... & Roussel, N. Technology probes: inspiring design for and with families. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 17-24). ACM (2003)

34. Jacquet, C., Bellik, Y., & Bourda, Y. Electronic locomotion aids for the blind: Towards more assistive systems. In Intelligent Paradigms for Assistive and Preventive Healthcare (pp. 133-163). Springer, Berlin, Heidelberg (2006)
35. Jafri, R., Ali, S. A., Arabnia, H. R., & Fatima, S. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. The Visual Computer, 30(11), 1197-1222 (2014)
36. Jayant, C., Acuario, C., Johnson, W., Hollier, J., & Ladner, R. V-braille: haptic braille perception using a touch-screen and vibration on mobile phones. In Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (pp. 295-296). ACM (2010)
37. Jayant, C., Ji, H., White, S., & Bigham, J. P. Supporting blind photography. In The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility (pp. 203-210). (2011)
38. Kane, S. K., Frey, B., & Wobbrock, J. O. Access lens: a gesture-based screen reader for real-world documents. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 347-350). ACM (2013)
39. Kawai, Y., & Tomita, F. A support system for visually impaired persons to understand three-dimensional visual information using acoustic interface. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 3, pp. 974-977). IEEE (2002)
40. Kayukawa, S., Higuchi, K., Guerreiro, J., Morishima, S., Sato, Y., Kitani, K., & Asakawa, C. BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 52). ACM. (2019)
41. Kelley, J. F. An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Office Information Systems, 2:1, 26–41 (1984)
42. KNFB Reader https://knfbreader.com/
43. Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105) (2012)
44. Kruger, C., & Cross, N. Solution driven versus problem driven design: strategies and outcomes. Design Studies, *27*(5), 527-548 (2006)
45. Kulyukin, V., & Kutiyanawala, A. 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, pp. 1-5) (2010)
46. Lanigan, P. E., Paulos, A. M., Williams, A. W., Rossi, D., & Narasimhan, P. Trinetra: Assistive Technologies for Grocery Shopping for the Blind. In ISWC (pp. 147-148) (2006)
47. LeCun, Y., Bengio, Y., & Hinton, G. Deep learning. nature, 521(7553), 436 (2015)
48. Lee, S., Reddie, M., Tsai, C., Beck, J., Rosson, M., & Carroll, J.M. The Emerging Professional Practice of Remote Sighted Assistance for People with Visual Impairments Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (HI, USA) (2020)
49. Lee, S., Yuan, C. W., Hanrahan, B. V., Rosson, M. B., & Carroll, J. M. 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 (pp. 389-390). ACM (2017)
50. Liu, X., Doermann, D., & Li, H. Mobile visual aid tools for users with visual impairments. In Mobile Multimedia Processing (pp. 21-36). Springer, Berlin, Heidelberg (2010)
51. Liu, M. Y., Lin, S., Ramalingam, S., & Tuzel, O. Layered interpretation of street view images. *arXiv preprint* (2015)

52. López-de-Ipiña, D., Lorido, T., & López, U. Blindshopping: enabling accessible shopping for visually impaired people through mobile technologies. In International Conference on Smart Homes and Health Telematics (pp. 266-270). Springer, Berlin, Heidelberg (2011)
53. Lowe, D. G. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110 (2004)
54. Makino, H., Ishii, I., & Nakashizuka, M. Development of navigation system for the blind using GPS and mobile phone combination. In Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE (Vol. 2, pp. 506-507). IEEE (1996)
55. Martinez, M., Roitberg, A., Koester, D., Stiefelhagen, R., & Schauerte, B. Using technology developed for autonomous cars to help navigate blind people. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1424-1432). (2017)
56. McDaniel, T. L., Kahol, K., Villanueva, D., & Panchanathan, S. Integration of RFID and computer vision for remote object perception for individuals who are blind. In Proceedings of the 2008 Ambi-Sys workshop on Haptic user interfaces in ambient media systems (p. 7). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2008)
57. Microsoft SoundScape https://www.microsoft.com/en-us/research/product/soundscape/
58. Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE transactions on robotics, *31*(5), 1147-1163. (2015)
59. Newell, A.F., Arnott, J.L., Carter, K. & Cruickshank, G. Listening typewriter simulation studies. International Journal of Man-Machine Studies, 33(1), Pages 1-19 (1990)
60. Oliva, A., & Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3), 145-175 (2001)
61. Persson, H., Åhman, H., Yngling, A.A. Gulliksen, J. Universal design, inclusive design, accessible design, design for all: different concepts—one goal? On the concept of accessibility—historical, methodological and philosophical aspects. Universal Access in the Information Society 14(4): 505-526 (2015)
62. Pfeiffer, D., Erbs, F., & Franke, U. Pixels, stixels, and objects. In European Conference on Computer Vision (pp. 1-10). Springer, Berlin, Heidelberg. (2012)
63. Pielot, M., Poppinga, B., Heuten, W., & Boll, S. A tactile compass for eyes-free pedestrian navigation. In IFIP Conference on Human-Computer Interaction (pp. 640-656). Springer, Berlin, Heidelberg (2011)
64. Rümelin, S., Rukzio, E., & Hardy, R. NaviRadar: a novel tactile information display for pedestrian navigation. In Proceedings of the 24th annual ACM symposium on User interface software and technology (pp. 293-302). ACM (2011).
65. Sánchez, J., & de la Torre, N. Autonomous navigation through the city for the blind. In Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (pp. 195-202). ACM (2010)
66. Scharwächter, T., Enzweiler, M., Franke, U., & Roth, S. Stixmantics: A medium-level model for real-time semantic scene understanding. In European Conference on Computer Vision (pp. 533-548). Springer, Cham. (2014)
67. Schauerte, B., Martinez, M., Constantinescu, A., & Stiefelhagen, R. An assistive vision system for the blind that helps find lost things. In International Conference on Computers for Handicapped Persons (pp. 566-572). Springer, Berlin, Heidelberg (2012)
68. SeeingAI https://www.microsoft.com/en-us/seeing-ai
69. Shilkrot, R., Huber, J., Meng Ee, W., Maes, P., & Nanayakkara, S. C. 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 (pp. 2363-2372). ACM (2015)
70. Siagian, C., & Itti, L. Rapid biologically-inspired scene classification using features shared with visual attention. IEEE transactions on pattern analysis and machine intelligence, 29(2), 300-312 (2007)

71. Stepnowski, A., Kamiński, Ł., & Demkowicz, J. Voice maps: the system for navigation of blind in urban area. In Proceedings of the 10th WSEAS international conference on Applied computer and applied computational science (pp. 201-206). World Scientific and Engineering Academy and Society (WSEAS) (2011)
72. Sudol, J., Dialameh, O., Blanchard, C., & Dorcey, T. Looktel—A comprehensive platform for computer-aided visual assistance. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 73-80). IEEE (2010)
73. Tanveer, M. I., Anam, A. S. M., Yeasin, M., & Khan, M. Do you see what I see?: designing a sensory substitution device to access non-verbal modes of communication. In Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (p. 10). ACM (2013)
74. TapTapSee https://taptapseeapp.com/
75. Tekin, E., & Coughlan, J. M. A mobile phone application enabling visually impaired users to find and read product barcodes. In International Conference on Computers for Handicapped Persons (pp. 290-295). Springer, Berlin, Heidelberg (2010)
76. Vergnieux, V., Macé, M. J. M., & Jouffrais, C. Wayfinding with Simulated Prosthetic Vision: Performance comparison with regular and structure-enhanced renderings. In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE (pp. 2585-2588). IEEE (2014)
77. Wang, D., Yang, Q., Abdul, A. and Lim, B.Y. Designing Theory-Driven User-Centric Explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland).(2019)
78. Wang, H. C., Katzschmann, R. K., Teng, S., Araki, B., Giarré, L., & Rus, D. Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. 2017 IEEE international conference on robotics and automation (ICRA) (pp. 6533-6540). IEEE. (2017)
79. Wang, J., Yang, K., Hu, W., & Wang, K. An environmental perception and navigational assistance system for visually impaired persons based on semantic stixels and sound interaction. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1921-1926). IEEE. (2018)
80. WebAIM. Screen Reader User Survey # 7 Results (2017)
81. Williamson, K., Wright, S., Schauder, D., & Bow, A. The Internet for the blind and visually impaired. Journal of Computer-Mediated Communication, 7(1), JCMC712 (2001)
82. Winlock, T., Christiansen, E., & Belongie, S. Toward real-time grocery detection for the visually impaired. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 49-56). IEEE (2010)
83. Yang, G., & Saniie, J. Indoor navigation for visually impaired using AR markers. 2017 IEEE International Conference on Electro Information Technology (EIT) (pp. 1-5). IEEE. (2017)
84. Yang, K., Bergasa, L. M., Romera, E., Sun, D., Wang, K., & Barea, R. Semantic perception of curbs beyond traversability for real-world navigation assistance systems. 2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES) (pp. 1-7). IEEE. (2018)
85. Yang, K., Wang, K., Cheng, R., Hu, W., Huang, X., & Bai, J. Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor. Sensors, *17*(8), 1890. (2017)
86. Yang, K., Wang, K., Chen, H., & Bai, J. Reducing the minimum range of a RGB-depth sensor to aid navigation in visually impaired individuals. Applied optics, *57*(11), 2809-2819. (2018)

87. Yatani, K., Banovic, N., & Truong, K. SpaceSense: representing geographical information to visually impaired people using spatial tactile feedback. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 415-424). ACM (2012)
88. Yuan, C.W., Hanrahan, B.V., Lee, S. & Carroll, J.M. Designing Equal Participation in Informal Learning for People with Visual Impairment. Interaction Design and Architecture(s) Journal - IxD&A, 27, 93-106 (2015)
89. Yuan, C.W., Hanrahan, B.V., Lee, S., Rosson, M.B. & Carroll, J.M. Constructing a holistic view of shopping with people with visual impairment: A participatory design approach. Universal Access in the Information Society (2017)
90. Yuan, C.W., Lee, S., Hanrahan, B.V., Rosson, M.B. & Carroll, J.M. I Didn't Know that You Knew I Knew: Collaborative Shopping Practices between People with Visual Impairment and People with Vision. Proceedings ACM Human-Computer Interaction, PACMHCI, Vol 1, Issue 2 (CSCW'18) (2018)
91. Yu, X., Yang, G., Jones, S., & Saniie, J. AR Marker Aided Obstacle Localization System for Assisting Visually Impaired. 2018 IEEE International Conference on Electro/Information Technology (EIT) (pp. 0271-0276). IEEE. (2018)
92. Zientara, P. A., Lee, S., Smith, G. H., Brenner, R., Itti, L., Rosson, M. B., ... & Narayanan, V. Third Eye: a shopping assistant for the visually impaired. Computer, 50(2), 16-24 (2017)