Iterative Design and Prototyping of Computer Vision Mediated Remote Sighted Assistance

JINGYI XIE, Pennsylvania State University, USA
MADISON REDDIE*, Pennsylvania State University, USA
SOOYEON LEE†, Pennsylvania State University, USA
SYED MASUM BILLAH, Pennsylvania State University, USA
ZIHAN ZHOU‡, Pennsylvania State University, USA
CHUN-HUA TSAI§, Pennsylvania State University, USA
JOHN M. CARROLL, Pennsylvania State University, USA

Remote sighted assistance (RSA) is an emerging navigational aid for people with visual impairments (PVI). Using scenario-based design to illustrate our ideas, we developed a prototype showcasing potential applications for computer vision to support RSA interactions. We reviewed the prototype demonstrating real-world navigation scenarios with an RSA expert, and then iteratively refined the prototype based on feedback. We reviewed the refined prototype with 12 RSA professionals to evaluate the desirability and feasibility of the prototyped computer vision concepts. The RSA expert and professionals were engaged by, and reacted insightfully and constructively to the proposed design ideas. We discuss what we learned about key resources, goals, and challenges of the RSA prosthetic practice through our iterative prototype review, as well as implications for the design of RSA systems and the integration of computer vision technologies into RSA.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *Accessibility*; Empirical studies in accessibility; Accessibility technologies.

Additional Key Words and Phrases: people with visual impairments, remote sighted assistance, computer vision, navigation, smartphone, augmented reality, 3D map

ACM Reference Format:
Jingyi Xie, Madison Reddie, Sooyeon Lee, Syed Masum Billah, Zihan Zhou, Chun-Hua Tsai, and John M. Carroll. 2021. Iterative Design and Prototyping of Computer Vision Mediated Remote Sighted Assistance. *ACM Trans. Comput.-Hum. Interact.* 1, 1, Article 1 (January 2021), 39 pages. https://doi.org/10.1145/3501298

*Also with Massachusetts Institute of Technology.
†Also with Rochester Institute of Technology.
‡Also with Manycore Tech Inc.
§Also with University of Nebraska at Omaha.

Authors' addresses: Jingyi Xie, Pennsylvania State University, University Park, Pennsylvania, USA, jzx5099@psu.edu; Madison Reddie, Pennsylvania State University, University Park, Pennsylvania, USA, mbr5511@psu.edu, reddie@mit.edu; Sooyeon Lee, Pennsylvania State University, University Park, Pennsylvania, USA, sul131@psu.edu, slics@rit.edu; Syed Masum Billah, Pennsylvania State University, University Park, Pennsylvania, USA, sbillah@psu.edu; Zihan Zhou, Pennsylvania State University, University Park, Pennsylvania, USA, zuz22@psu.edu; Chun-Hua Tsai, Pennsylvania State University, University Park, Pennsylvania, USA, ctsai@psu.edu, chunhuatsai@unomaha.edu; John M. Carroll, Pennsylvania State University, University Park, Pennsylvania, USA, jmc56@psu.edu.

1 INTRODUCTION

People with visual impairments are increasingly relying on remote sighted assistants for indoor or outdoor navigation [88]. This form of assistance is generally called *remote sighted assistance (RSA),* a service (e.g., Be My Eyes [4], Aira [1]) in which people with visual impairments (PVI) establish a video connection with a remote sighted assistant (namely, *RSA agent* or simply *agent*), who then interprets the video stream coming from the PVI's smartphone camera, while conversing with them to provide assistance as needed or requested.

Early RSA services [35, 56, 84, 109] were largely characterized by unidirectional communication, (e.g., agent to PVI) and narrow scope (e.g., focused only on navigation or identifying objects in a single static image). Over time, these services adopted new technologies and broadened their scope. For instance, in the current generation of RSA services (e.g., Aira [1]), agents leverage a variety of technologies, including pervasive networking infrastructures to create two-way, audio/video communication; location services (e.g., GPS); mapping services (e.g., Google Maps); and information technology (e.g., Google Search) [88]. Furthermore, with the assistance of remote sighted agents, PVI can now perform a set of broader and more complex tasks, such as navigating airports and shopping in large malls [88], which usually required in-person sighted assistance in the past.

As RSA services have broadened in scope, current RSA technologies have become the limiting factor for the agents, impacting their performance and subsequently degrading the overall quality of service experienced by PVI. For example, agents face the following challenges in video-mediated RSA services during their interactions with PVI [77, 88]: (i) they lack confidence due to unfamiliarity of the PVI's current physical environment; (ii) do not often have indoor maps with fine details; (iii) need to track the PVI continuously on maps and orient them within their surroundings manually; (iv) need to estimate objects' depth in the video stream coming from the PVI's camera and describe those in real-time; (v) need to detect landmarks visually and track dynamic objects mentally; and (vi) need to develop mutual trust and synergy with PVI.

Fortunately, a subset of these challenges are well studied in computer vision (CV) and AI research under the categories of indoor map construction and localization [37], depth estimation [45, 124], object tracking [131], visual navigation [28, 163], object recognition and scene understanding [67, 104, 114], and explainable AI [63, 115]. However, problems in the above categories are generally considered hard, and current solutions are not reliable enough to deploy for people with visual impairments [120].

Drawing on these rich bodies of literature, we propose to investigate whether RSA agents can adopt CV technology, complementing their existing technologies, in order to address the aforementioned challenges. More specifically, we investigate a model of CV-mediated RSA service, where RSA agents are the immediate users of the CV technology. Our goal is to explore whether integrating CV can produce a new, desirable experience for the agents so that they can provide better quality service to PVI.

Recent prior work has investigated the perspective of PVI with regard to RSA services [42, 77, 87], and RSA services have traditionally been developed based on PVI's needs and feedback after trials by PVI [24, 35, 84, 109, 126]. However, we argue that the agents' perspective is increasingly important in designing new RSA systems that depend on their performance. In addition, RSA agents are experienced professionals who have assisted numerous PVI with diverse preferences and in diverse contexts and activities [88].

We first identified potential applications for CV to assist RSA agents and proposed design ideas regarding how to incorporate CV into current RSA interactions. Next, we embedded those design ideas into different real-world scenarios. For each scenario, we developed a series of low-fidelity prototypes with static images, animations, and narrations. We

presented these prototypes to a domain expert as probes and then iteratively refined the prototypes based on feedback. We reviewed the refined prototypes with 12 RSA agents (i.e., trained professionals) to evaluate the desirability and feasibility of the prototyped CV concepts and system designs.

Our analysis shows that all of our design ideas with CV concepts have promising potential to alleviate major navigational challenges and to enhance the RSA assistive experience. Our findings suggest that the design ideas proposed will help augment and extend the agents' vision in different dimensions, which gives them the ability to see further spatially and predictably as well as holistically for them to stay ahead and to manage possible risks. Agents also richly describe how helpful and useful each idea is for mitigating the challenges in their everyday work. Lastly, opportunities for improvement in the concepts and prototypes are identified.

We note that PVI are beneficiaries of and central stakeholders in CV applications in RSA services, and as such, should be engaged in the design process. This is, however, beyond the scope of this paper, which is a first step in exploring possibilities for CV to assist RSA agents.

In summary, the contributions of this work are as follows:

- We adopted a scenario-based design to develop low-fidelity prototypes illustrating how computer vision can help remote sighted assistants, who assist blind users in navigation and other tasks remotely.
- We conducted semi-structured interviews with professional remote sighted assistants to evaluate the desirability and feasibility of our prototyped computer vision concepts, as well as discovered key themes that can be addressed by computer vision and related technologies.

2 BACKGROUND AND RELATED WORK

Researchers have proposed many prototypes to aid people with visual impairments in outdoor and indoor navigation [112]. They identified that such navigational aids must have two components to facilitate independent mobility: (i) *obstacle avoidance*, and (ii) *wayfinding* [111]. Obstacle avoidance ensures that visually impaired users can move through space safely without running into objects. Guide dogs and white canes are often used for this purpose. Wayfinding, on the other hand, allows them to plan and execute a route to a desired destination. For wayfinding, having a representation of users' surroundings (i.e., digital maps, cognitive maps [142], building layouts) is essential, and so is *localization*, (i.e., continuously updating their location within that representation).

Recently, smartphone-based wayfinding apps have become mainstream for outdoor navigation for people with and without visual impairments. These apps, such as Google Maps [6], BlindSquare [27], SeeingEyeGPS [13], Soundscape [134], and Autour [3], rely on GPS for localization and commercial map services (e.g., Google Maps Platform [16], OpenStreet Map [17]) for wayfinding.

Although blind and visually impaired users can navigate large outdoor distances using these apps, they struggle to find the last-few-meters [120] due to a wide margin of error in GPS accuracy (±5m [61]). They struggle even more during indoor navigation because of the weaker GPS signal strength in the indoor environment that renders these apps unreliable, and the lack of sufficiently detailed indoor map data [90, 117].

To overcome these limitations, researchers have proposed fusing GPS signal with smartphones' built-in sensors, such as motion sensors, Bluetooth [122], Infrared [89], NFC [55], RFID [54], sonar [44], and camera. Also, they have made a concerted effort to construct indoor maps and extract the semantic features of the environment [47]. Unfortunately, these solutions require additional deployment and maintenance effort to augment the physical environment as well as significant bootstrapping costs for setting up databases of floorplans [49] and structural landmarks [22, 108]. Some

solutions also require users to carry specialized devices (e.g., an IR tag reader [89]). For these reasons, no single indoor navigation system is widely deployed. In this work, we envision that the recent deployment of smartphone-based augmented reality frameworks and popular human-assisted remote navigation services could pave the way for constructing indoor maps organically.

2.1 Remote Sighted Assistance Services for People with Visual Impairments

The concept of the remote sighted assistance (RSA) service has evolved over time, from the early idea of tele-assistance using information and telecommunication technology, to crowdsourced assistance using smartphone applications, to paid assistance using smartphones and specialized hardware.

In early prototypes of RSA services, the remote sighted assistants and the blind users communicated via synthetic speech [109], images [26, 84], one-way video using portable digital cameras [35, 56], and webcams [35]; whereas in recent ones like [1, 4, 24, 70], they are using two-way video chat with smartphones. In addition, the localization technique used in these services progressed continuously – starting from GPS-only, later augmented by fusing sensors [111], crowdsourcing [26, 85, 111, 160], and CV [34].

Several researchers examined the feasibility of crowdsourced RSA services (e.g., TapTapSee [14], BeMyEyes [4]) and concluded that this is a promising direction to tackle navigation challenges for blind users [21, 29]. Burton et al. [36] studied how crowdworkers answer subjective questions asked by blind participants. They commented on the issue of blind users trusting the responses of sighted crowdworkers too much, even though some crowdworkers are not experts, and many are not available at times. Nguyen et al. [103] and Lee at al. [88] studied a paid RSA service, Aira [1]. They reported that unlike crowdworkers, Aira agents are always available and trained in communication terminology and etiquette. Furthermore, they do not provide subjective information. In this paper, we assume that Aira or a similar RSA service exists to demonstrate our design.

2.2 Computer Vision Capabilities

The past decade has witnessed the rapid development and commercialization of CV technologies, thanks to the availability of large-scale visual data and the emergence of new computational tools (e.g., deep learning). In this section, we review the CV technologies used in our design and prototype. Most of them can be classified into the following categories and are en route to maturity in coming years.

2.2.1 Mature Technologies. Technologies such as real-time object detection and scene text reading are now readily available as commercial products, including those designed to assist PVI (e.g., Microsoft SeeingAI [101]).

Object detection can be categorized into two types [62, 91, 159]: identifying specific instances (e.g., famous paintings and landmarks) and generic object categories (e.g, cat and dog). Both problems have been addressed successfully. Recently, deep learning methods have increased accuracy dramatically, especially in the domain of generic object detection [59, 60, 83, 113].

Text recognition has been applied to help PVI for decades, including reading currency [92, 106, 107], signs [99, 128], and document text [48, 78, 129]. Like object detection, deep learning methods have improved the accuracy of text recognition, by addressing the challenges of text rotation and perspective changes [68, 96, 162], densely arranged text detection [93, 95, 152], and broken and blurred text detection [73, 74, 140].

2.2.2 Emerging Technologies. Technologies such as pedestrian detection, path tracking and prediction, and obstacle distance estimation play critical roles in emerging applications (e.g., autonomous driving [38], driver assist [58, 76]). These exist now and are improving rapidly in reliability, stability, and efficiency.

Deep learning has greatly promoted the progress of pedestrian detection [165]. Feature fusion [156] addressed the problem of detecting small pedestrians. The integration of boosted decision tree [156] and semantics segmentation [139] is a recent solution to improve hard negative detection. Other methods [105, 138, 148, 158] have improved dense and occluded pedestrian detection.

Path prediction needs to extract more information from video than recognition tasks, including the information of the surrounding environment and status of prediction targets [69]. Researchers have used semantic segmentation [23, 79, 86] and convolutional neural network (CNN) [71, 133] to retrieve environmental features. Target features include the orientation of the target [71, 81, 97] and physical attributes (e.g., age and gender). Based on the feature extraction from video, different methods have been applied for path prediction, such as Bayesian models [23, 81, 127], energy minimization [71, 145, 150], and deep learning [19, 50, 86, 153].

Research has been conducted to estimate non-contact distance of objects within the camera's field of view. Some methods utilize extra rangefinders, which align laser rangefinders with cameras [57, 125, 157]. Researchers have also used binocular stereo vision setups to track the locations of objects, which can estimate distance in 3D space [141, 151]. Another approach is to calculate the distance through CCD cameras [94, 147].

2.2.3 Use of Computer Vision in Navigation for People with Visual Impairments. Budrionis et al. [34] reported that CV-based navigation apps running on smartphones are a cost-effective solution for indoor navigation. A primary focus in the CV-based approach is how to make visual information more accessible through recognizing objects [164], color-codes, and landmarks (e.g., storefronts [120], signage [51]), and through the processing of tags such as barcodes, RFID, or vanishing points [46, 100, 137]. Extending this focus, researchers have proposed several indoor positioning and navigation systems [80, 89, 98]. However, Saha et al. [120], who studied the last-few-meters wayfinding challenge for people with visual impairments, concluded that for a deployable level of accuracy, using CV techniques alone is not sufficient yet. Our work aligns with the findings of [120]; we propose using CV techniques to assist remote sighted assistants (e.g., RSA agents), rather than people with visual impairments, who are vulnerable to inaccuracies.

Lately, researchers are exploring the potential of augmented reality (AR) frameworks in indoor navigation. These frameworks are built into modern smartphones (e.g., ARKit [8] in iOS devices, ARCore [2] in Android devices), and thus have the potential for widespread deployment [116]. Based on ARKit, Verma et al. [143] proposed an indoor navigation application and reported that an AR-based navigation system could provide a better user experience than traditional 2D maps. Clew [154] demonstrated the potential of constructing indoor 3D maps using ARKit and localizing blind users on that map with acceptable accuracy. Fusco et al. [51] also reported that with ARKit, users do not need to aim the camera towards an object to recognize it, which could be convenient for visually impaired users. In this paper, we explore challenges in utilizing AR frameworks in RSA systems to benefit remote sighted assistants.

2.3 Collaboration between Humans and AI

Automatic scene understanding from video streams and 3D reconstruction remain challenging despite recent CV advancements. Factors, such as motion blur, image resolution, noise, change of light, scale, and orientation, impact the performance and accuracy of existing systems [75]. To overcome these challenges, researchers have proposed interactive, hybrid approaches that involve human-AI collaboration [30]. One variation of hybrid approaches is the

human-in-the-loop framework. Branson et al.'s [31] system utilized human responses to questions posed by the computer to drive up the recognition accuracy while minimizing human effort. Similarly, some researchers developed interactive 3D modeling in which humans draw simple outlines or scribbles to guide the process [82, 130]. In this work, we developed a series of low-fidelity prototypes to understand the challenges in human-AI interaction in RSA services.

2.4 Low-Fidelity Prototyping

Prototyping refers to the development of partial and/or tentative implementations of a system design. The key motivation for prototyping is to make it more possible to analyze and assess designs without first incurring the costs and the work of fully implementing the designs. This concern has encouraged the development of a wide range of low-fidelity prototyping methods, such as paper prototyping [132], where the layout and key interactions of a user interface are mocked up with bits of paper, and Wizard of Oz performative prototyping [39], where a (concealed) human plays the part of an intelligent agent or other interactive capability.

One strength of low-fidelity prototypes is that they can be implemented relatively quickly and inexpensively to help designers reflect more concretely on a design, or even to evoke user experiences and reactions to the design. A potential downside of low-fidelity prototyping is that the prototype may be too crude to evoke experiences in designers and potential users that are useful in assessing and further developing the design. For example, one would not want to use a low-fidelity prototype to investigate temporal parameters for rapid input-output interactions. However, low-fidelity methods have a wide range of fundamental application and user interface issues [41, 123, 144].

2.5 Crowdsourcing Map Construction

Crowdsourcing has become technologically and logistically possible, and even routine, for 2D map construction and maintenance. OpenStreetMap [11], an open-source and open-access project founded in 2004, is widely considered to be the pinnacle of volunteer crowdsourced map construction [25]. Apart from map-making, OpenStreetMap contributors can annotate maps by adding labels (e.g., environmental tags) to spatial features. In addition to crowdsourcing of outdoor mapping, the ubiquity of mobile devices (i.e., smartphones) has facilitated crowdsourced indoor mapping [161]. Researchers have developed several systems for crowdsourcing-based indoor map construction, such as CrowdInside [20], SAMS [110], and CrowdMap [43]. The success of these projects attests to the feasibility of crowdsourcing-based 2D map construction for both outdoor and indoor environments and suggests that our design idea of crowdsourced 3D map construction may likewise be promising.

Researchers have also probed the use of crowdsourcing in supporting navigational tasks for PVI, including improving public transit accessibility [65] and providing rich information about intersection geometry [64]. Hara et al. [65] found that judging landmarks' proximity to a target object (a bus stop) from a static image is hard, which leads to mislabeling or over-labeling. They suggested that using 3D maps is more reliable for estimating physical placement compared with the 2D imagery used in their study. Guy and Truong [64] indicated that crowdsourced annotations represent information requested by PVI users and compensate for information not available in current open databases. This work paved the path for crowdsourcing-based navigation aids for PVI and supported the rationality of some of the proposed design ideas in this study.

Although prior work supports the technological feasibility of crowdsourced mapping, the motivation and incentives of volunteers have been a concern surrounding crowdsourced map construction. In the case of OpenStreetMap, researchers have investigated 39 motivations for contributory behavior in open collaboration [33]. Reciprocity, altruism, the instrumentality of local knowledge, social relations, and self-actualization are some motivators for volunteers to upload

geographic information and edit and annotate maps. Businesses may be incentivized to map their locations if their accessibility makes them attractive to visually impaired consumers. Many other types of organizations, and even local governments, also value accessibility either formally or informally, which could motivate them to map indoor facilities.

3 DESIGN PROTOTYPE

In our previous work with RSA agents and PVI [88, 155], we identified a number of challenging navigation scenarios for both agents and PVI. Such scenarios include navigating airports [40, 88], grocery shopping [155], navigating malls or buildings with coarse or poor maps [88], walking down crowded streets [88], and walking in parks [40]. In Section 3.1, we summarize the types of challenges that are characteristic of these difficult scenarios, according to the literature. We then describe how computer vision-inspired design ideas can potentially benefit RSA agents in scenarios affected by these challenges based on our prior work in [40]. Finally, we describe our method to contextualize the design ideas in different real-world scenarios and iteratively refine our prototype based on review feedback in Section 3.2.

3.1 Navigation Challenges and the Envisioned Design Space

As a first step toward developing CV-mediated remote-sighted assistance, we identified eight root causes from the literature that can make a navigation scenario challenging to RSA agents. These causes or challenges include: C1) *Lack of indoor maps*: Unlike outdoor navigation, which can rely on open maps and GPS, indoor navigation is usually lacking the proper support of an indoor map [77, 88, 102, 111]. The RSA agent needs to either read the building layout through the camera or rely on the PVI's feedback; C2) *Localizing the PVI on the map in real-time*: Even equipped with a map, the agent needs to localize the PVI, which takes considerable, constant effort [24, 56, 88, 102]; C3) *Orienting the PVI in their current surroundings:* The agent needs to become familiar with the surroundings quickly to provide real-time navigation to the PVI. Environmental orientation is difficult and often time-consuming [35, 56, 111]; C4) *Estimating depth from the PVI's camera feed:* It is challenging to estimate the actual distance of an object (e.g., to tell the PVI how far away doors are), due to the variable quality, angle, and stability of the video feed [77]; C5) *Reading signage and text in the PVI's camera feed:* Sometimes, the agent needs to read signage or text though the PVI's video feed; it is difficult to read text in hand-held shot video [70]; C6) *Detecting and tracking moving objects:* Moving objects, e.g., vehicles in traffic or pedestrians on the sidewalk, are hard to detect and track because the PVI needs to hold the phone and follow the object, which is not realistic [70, 72]; C7) *Projecting or estimating out-of-frame (blocked) objects from the PVI's camera feed:* If objects are out of the video feed, e.g., the signs or items just passed the PVI, they are difficult to keep track of [24, 35, 56, 72, 77, 109, 126]; C8) *Unstable network connection:* Agents rely on the real-time video feed to receive necessary information from the PVI, but it is a challenge if a stable connection cannot be established [42, 56, 70, 72, 77, 88].

ID	Design Idea	Challenges Addressed
D1	3D map construction	C1, C2, C3
D2	Augmenting existing 2D maps	C1, C2, C3
D3	Interacting with 2D/3D maps	C1, C2, C3
D4	Augmenting video stream	C3, C4, C5, C8
D5	Mapping and navigating dynamic scenes	C6, C7

Table 1. A list of design ideas for addressing the navigation challenges.

To address these challenges, we define our design space as five main design ideas, as summarized in Table 1. The "map" component is essential in multiple challenges above, especially in indoor navigation (C1), localization (C2), and environmental orientation (C3). In our design space, we have three design ideas aimed at improving maps for navigation, including 3D map construction (D1), augmenting existing 2D maps (D2), and enabling interaction with 2D/3D maps (D3).

3D maps (D1) profile real-world objects in 3D space, which is represented by point cloud data. We can currently use smart devices to generate point cloud data with built-in AR frameworks (e.g., ARKit [8] in iOS devices, ARCore [2] in Android devices). Regardless of the layouts of different buildings, 3D maps can be constructed in the same way: volunteers collect, upload, and update geographical data collaboratively via the Internet. This approach is an application of crowdsourcing and detailed procedures are described as follows. Sighted volunteers can scan areas with mobile devices to generate point cloud data that is uploaded to servers in real-time, and mappings of the same location can be merged from information provided by different volunteers. Likewise, sighted volunteers and RSA agents could annotate 3D maps by creating, editing, and deleting labels. In the absence of financing to construct and maintain 3D maps, crowdsourcing these tasks could optimally make relatively up-to-date indoor maps available and address C1.

When a PVI enters a building, the offline-built 3D map of the space (if available) will be loaded automatically to the RSA agent's dashboard. To enable continuous tracking of the PVI in real-time, feature points can be detected from the frames of the live camera feed and matched with the 3D point cloud. We assume that the PVI holds the smart device in a front-facing manner, so the PVI's location (C2) and orientation (C3) are the same as that of the camera.

In some situations, a 2D map is available, such as Google Maps [6] for outdoor navigation or static, offline airport maps. We use CV to augment these maps (D2) so that RSA agents can use them to better assist PVI. For example, object detection technology can find relevant objects (e.g., trash cans, benches, vending stands) in a live video feed. We would store the objects in the map, and the RSA agents could retrieve them later, when needed. As another example, CV can recognize zebra crossings to help RSA agents find a safe path to a PVI's destination, and this data can be added to maps and reused in other RSA interactions in the same location.

Interacting with 2D/3D maps (D3) will allow RSA agents to change the scale and orientation (e.g., via zoom in/out and rotation), which can help with localization (C2) and orientation (C3). The envisioned map interaction feature also allows RSA agents to manually draw and edit planned navigation paths on the maps to facilitate efficient and safe navigation.

We will also enhance and enrich the video stream by integrating information from maps and using CV recognition features (D4). For instance, we can leverage data from maps to present key, real-time turn-by-turn directions (C3) and distances (C4), contextualized within the video feed, with AR technologies. Enhanced text overlaid on the text in the video recognized and read by CV could alleviate the difficulties RSA agents have reading small, rotated text in live camera feeds (C5), even when frames are corrupted due to unreliable connections (C8).

Our final design idea is to interpret dynamic scenes to address the challenge of dynamic and out-of-frame objects (D5). We can use pedestrian trajectory forecasting to avoid potential collisions with moving objects (C6) and predict the future motion of a person who disappears from the camera's field of view (C7).

3.2 Method

We adopted scenario-based design in this paper and contextualized our design ideas in five real-world scenarios. Scenario-based design is to *use a future system concretely described at an early point in the development process. Narrative descriptions of envisioned usage episodes are then employed in a variety of ways to guide the development of the system*

Scenario	Instantiation of Design Ideas before Expert Review	Instantiation of Design Ideas after Expert Review
Walk in the park	D2: use compass to orient the PVI; D4: estimate distance and show distance bands on ground; D5: continuous pedestrian tracking and motion forecasting on the map; show movement predictions in video feed.	D2: use compass to orient the PVI; *detect relevant objects in live video feed and store them in the map;* D3: *provide search bar to look up relevant objects;* D4: estimate distance and show distance bands on ground; *detect and highlight obstacles at least 30ft away*; D5: continuous pedestrian tracking and motion forecasting on the map; *identify distracted people*; show movement predictions in video feed.
Airport navigation	D3: use the indoor location and mapping service available at an airport (e.g., LocusMaps [7]) to track the PVI and wayfind; D4: recognize landmarks (e.g., monitors, moving walkways) and signage, read scene text (e.g., "EXIT"); show distance bands on ground; show distance to objects; show walking directions; D5: detect obstacles and crowds; staff recognition.	D3: *agents draw and highlight motion path/plan on 2D maps; continuously track the PVI on the path/plan;* D4: recognize landmarks (e.g., monitors, moving walkway) and signage, read scene text (e.g., "EXIT"); show distance bands on ground; show distance to object marks; show walking directions; D5: *detect obstacles and queues;* staff recognition.
Office building navigation	D1: construct and label indoor 3D maps collaboratively; D3: show the PVI's location on the map continuously; zoom in/out and rotate 3D maps; double-click to toggle views; first-person views (side-by-side or overlay display); plan path (manually or automatically); D4: show walking directions; show distance bands on ground; recognize objects and landmarks.	D1: construct and label indoor 3D maps collaboratively; D3: show the PVI's location on the map continuously; zoom in/out and rotate 3D maps; double-click to toggle views; first-person views (side-by-side or overlay display); plan path (manually or automatically); D4: show walking directions; show distance bands on ground; recognize objects and location marks.
Navigate from parking lot	D2: recognize cues close to entrance (e.g., zebra crossing, accessible parking, logos/signs); line up indoor map and satellite image on Google Maps; D4: show walking directions.	D2: recognize cues close to entrance (e.g., zebra crossing, accessible parking, logos/signs); line up indoor map and satellite image on Google Maps; D4: show walking directions.
Find rideshare	D4: project the pickup vehicle from the rideshare application map to the video feed.	
Grocery store shopping		D1: *construct 3D maps of store structures (e.g., sections, aisles);* D2: *detect relevant objects in live video feed and store them in the map;* D4: *recognize landmarks and read scene text; localize the PVI under unstable network connection using AR points; augment video feed with first-person view of 3D map.*

Table 2. Scenario-based design for expert review and design review study. Items in *Italics* were developed based on the expert review.

that will enable these use experiences [119]. Our goal is to provide narrative descriptions that can be adopted and used in future implementations. This method is ideal when the user need is not well-defined or when adopting new technology in an existing context, e.g., AI-enabled RSA navigation for PVI.

Based on the literature and our experience working with PVI and RSA agents, we developed several navigation scenarios, which we narrowed down to five. One of our initial scenarios (find rideshare) was replaced by another deemed more relevant and practical (grocery store) through discussion during the expert review, described in more detail later. The resulting five scenarios are important, frequent, and challenging in RSA practice; capture a diversity of contexts; and allow the insertion of CV technology in ways that we hypothesized that agents would experience as valid and engaging. In Table 2, we define the name of each scenario and the instantiations of the design ideas. We proposed low-fidelity prototypes for each scenario, which can be found in Figures 1 and 2.

In this paper, we adopted *Aira RSA Service* as our research platform. Aira [1] is a commercially available, on-demand RSA subscription service for PVI. As of 2019, Aira has offered service to thousands of "explorers" (a term that they use to refer to their users), advancing their learning, performance, and employment opportunities. We conducted two studies to collect feedback from the expert and the end-users (RSA agents). All interviews were conducted over Zoom (recorded, after consent) and lasted for an hour. We first described a scenario setting in words and asked participants (RSA agents) whether they had assisted any PVI in that circumstance. All participants had indeed had experience with all five scenario types. For each scenario, we then presented our prototypes in PowerPoint slides via screen sharing. All interviews were transcribed by two researchers.

We used a bottom-up approach in our qualitative data analysis. Two researchers independently performed inductive thematic analysis [32] on the transcribed interviews. Both of them used open coding to develop initial codes and generated categories and subthemes through iterative collating and grouping. The categories and subthemes generated by each researcher were reviewed and finalized in meetings with all authors. The agreed-upon subthemes shown in Table 3 from both studies 1 and 2 were further organized into the following overarching five themes: reducing the agent's cognitive load, enhancing the agent's ability to stay ahead, contextualizing object detection, emphasizing the PVI's video feed, and managing risk in navigation (as well as managing external factors in study 1 only) (Table 3).

3.2.1 Study 1. In study 1, we interviewed a domain expert, who is the most senior staff member in managing the RSA service and platform at Aira. The first five scenarios in Table 2 were presented to this participant (P0) in the order shown in the table. Once all five scenarios were presented, P0 shared her feedback in an unstructured interview, with five researchers asking follow-up questions and providing clarifications. This study contained 16 (=3 of the first scenario +3 of the second +7 of the third +2 of the fourth +1 of the fifth) illustrations in total. P0's feedback was used to improve our design ideas and prototypes for study 2.

3.2.2 Study 2. In study 2, we showed the improved design ideas and prototypes to the end-users, the RSA agents. Professional, trained RSA agent volunteers (not compensated for their participation) were recruited by Aira on our behalf as a part of a continuous research partnership with Aira. The expert in study 1 (P0) helped us to send out invitation emails and recruit the agents. Only RSA agents with extensive RSA experience (greater than 1 year) were eligible to participate. A total of 12 agents (P1-P12) participated in this study. The demographic information of the agents can be found in Table 4. We obtained IRB approval from our institution for the human participants.

Interviews with P1-P12 were semi-structured, with between two and five researchers in each. The five revised scenarios resulting from the design iteration were presented to P1-P12 in the order shown in Table 2. The final iteration contained 18 (=4+3+8+2+1) illustrations in total.

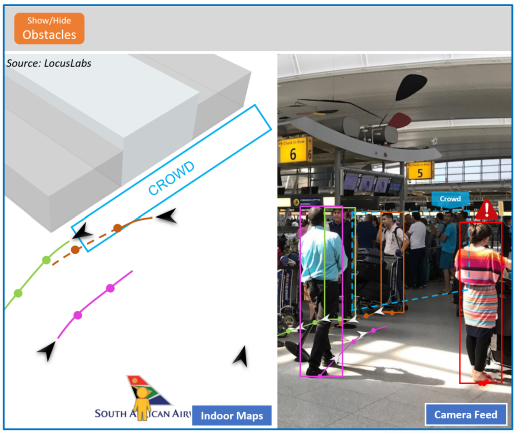
Themes	Subthemes
Reducing the Agent's Cognitive Load	(1) Orientation (2) Reducing Reliance on Scanning (3) Translation and Conversion
Enhancing the Agent's Ability to Stay Ahead	(1) Distance Estimation (2) Path Planning and 3D Maps (3) Direction Tracking and Prediction
Contextualizing Object Detection	(1) Selective Pedestrian, Object, and Obstacle Detection (2) Scene Text Reading (3) Staff Recognition (4) Toggling Features On and Off
Emphasizing the PVI's Video Feed	
Managing Risk in Navigation	(1) Parking Lots (2) Applications to the COVID-19 Pandemic (3) Unstable Network Connection (4) Further Opportunities
Managing External Factors	

Table 3. Themes and subthemes

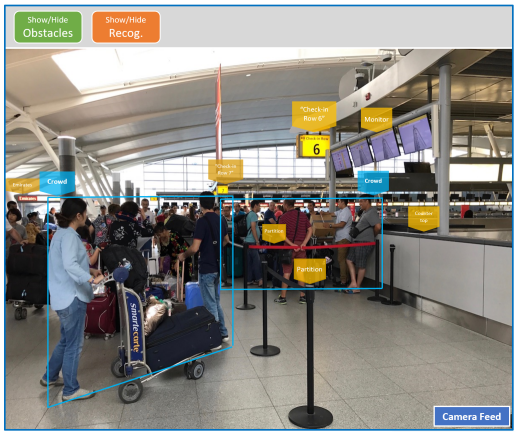
At any time during a scenario walkthrough, participants were allowed to ask questions, make comments, or request to revisit previous slides. We also encouraged them to express their thoughts and feedback before moving on to the next scenario. Depending on the amount of feedback given during a walkthrough, researchers then asked about participants' current methods of coping with the problems described in the scenario, and whether the prototypes addressed all of the challenges associated with the scenario type. If time remained in the hour after the discussion of the fifth scenario, researchers asked follow-up questions, questions about general features that appeared in multiple scenarios (e.g., the interface presented), if there were any other navigation challenges not covered by the five scenarios, what participants' favorite features were, and if any features appeared problematic or could be improved.

4 STUDY 1: EXPERT REVIEW

Our interview with the domain expert (P0) validated nearly all of the design ideas that we presented and revealed several themes describing the relationship between RSA practice and our design ideas: reducing the agent's cognitive load, enhancing the agent's ability to stay ahead, contextualizing object detection, emphasizing the PVI's video feed, managing risk in navigation, and managing external factors. These themes illuminate why the design ideas could be

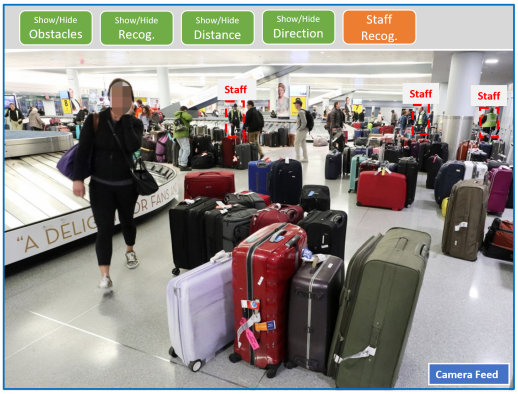


(a) Scenario 2: crowd recognition & pedestrian tracking



(b) Scenario 2: landmark recognition & scene text reading

(c) Scenario 2: indoor localization & AR directions



(d) Scenario 2: staff recognition

Fig. 1. Our design prototype for the "airport navigation" scenario. (a) PVI is approaching a check-in desk; (b) PVI is waiting in line to check-in; (c) PVI is navigated to a security check, (d) PVI is finding luggage in a baggage claim area. The top toolbar in each figure shows buttons to toggle a design feature on or off. The information on indoor maps and the camera feed is coordinated through color. Rectangles represent pedestrian detection; lines on the ground are trajectory predictions; intervals between dots symbolize equal distance; arrows represent orientation; alerts will pop up when collisions may occur; and blue chevron arrows represent planned path. Sample images used in the interface are originally drawn from the following sites (top-left, clockwise): locuslabs.com [9], onemileatatime.com [10], upi.com [15], thepointsguy.com [12], and locuslabs.com [9].

beneficial, and the agent values, goals, and challenges that P0 uses to evaluate the ideas provide further insight into how the concepts can be optimized and other ways that we may supplement RSA.

According to our findings from the expert review, we further improved the design ideas. The changes made with respect to each theme are described in the "Design Iteration" subsection of each theme. Table 2 summarizes the revised designs.

(a) Scenario 1: walking in the park

(b) Scenario 3: office building navigation

(c) Scenario 4: navigating from parking lot

(d) Scenario 5: find rideshare

Fig. 2. Examples of our design prototypes for the other four scenarios. The top toolbar in each figure shows buttons to toggle a design feature on or off. The information on indoor maps and the camera feed is coordinated through color. Representations of *pedestrian detection and tracking*: rectangles represent pedestrian detection; solid lines on the ground mean trajectory prediction; intervals between dots symbolize equal distance; arrows represent pedestrians' orientation; alerts will pop up when collisions may occur. Representations of *distance measurements and path planning*: dashed lines on the ground represent distances to obstacles, chevron arrows represent planned paths. Representations of *AR tracking*: the cube symbolizes the car, arrows and dashed lines represent predicted orientation and trajectory. Sample images used in Figure 2c interface are originally drawn from maps.google.com [5].

4.1 Reducing the Agent's Cognitive Load

During navigation tasks, agents must continually absorb and process multiple streams of dynamic information, including obstacles in the PVI's immediate path, other scenery of interest in the video feed, the PVI's location on maps, verbal and nonverbal communication from the PVI, and any additional information that the PVI has requested. As a result,

ID	Gender	Age	Time as an agent	Work hours (Avg.)	Device use	Background/Occupation
P1	M	56	1 yr	4-6 hrs a week	N/A	military, law enforcement, call center
P2	F	36	3 yrs	4-5 days a week, 7-8 hrs a day	laptop	dispatcher, paralegal, bartender
P3	F	27	3.5 yrs	previously 40-60 hrs, now 2 hrs a week	PC	Ph.D. occupational therapy
P4	M	31	3 yrs	40 hrs a week	laptop (Mac), 2 monitors	health sciences, in-home healthcare, sales
P5	M	30	1 yr	40 hrs a week	PC, single screen	“different kinds of work," likes helping people
P6	M	31	1 yr	30-40 hrs a week	PC, 3 monitors	languages, political science, non-profits
P7	F	28	14 mos	about 40 hrs a week	laptop, 2 monitors	medical technician
P8	F	28	2.5 yrs	40 hrs a week	laptop (Mac), 2 monitors	Spanish studies, relations manager
P9	F	25	3 yrs	25 hrs a week	laptop (Mac), single screen	kinesiology, first job
P10	F	28	1.5 yrs	20 hrs a week	laptop, 2 monitors	psychology, HR
P11	N/A	31	3 yrs	26-29 hrs a week	PC and Mac, 3 monitors	global studies, English, Japanese, health office worker
P12	F	44	3 yrs	4-8 hrs a week	laptop, monitor	elementary education

Table 4. Demographic information of Aira agents participated in Study 2.

the agent’s cognitive load can be quite high, and agents may struggle to keep up with the constant flow of incoming information. P0 explained one example of how she must pay attention to the PVI while determining their location, searching the internet for information, and path planning.

> *“... while [the PVI is] telling you everything about their day and how they got to the airport and what it is they’re doing, the agent is actively working to find which airport they’re at... and start route planning... I’m over on the side, like, frantically looking up the internal airport map.”*

We can see that managing high cognitive load can be difficult and stressful. P0 appreciated how some of the design ideas could reduce agent cognitive load. For example, she liked how the room annotations in the office building scenario (Figure 2b) would eliminate the need for her to look at a map to find a certain office.

> *“I really like the annotated room names... People regularly go to the same place day in, day out... I just imagined being a brand new Aira agent who had never been to somebody’s office before, and maybe there’s a map that is marked as, like, ‘my office,’ and so I could just simply go right in and lead that person there...”*

P0 also suggested revisions to the features where she perceived high potential cognitive load. For example, in the airport navigation scenario, we showed a map with each airline’s check-in area annotated (Figure 1b). Rather than labeling specific airline check-in desks on airport maps, P0 suggested that the general check-in area be marked and then the text-reading feature could be employed to help agents find the correct desk to reduce the amount of information the agent must process at once.

> *"If I know where the check-in desks are, I can get closer to them, and then from there, if you were able to incorporate the text reading, then we'd be able to kind of fill in some of those gaps one step at a time versus having that information right away."*

Similarly, she commented that we should limit the number of toggle buttons (Figure 2) that we include so as not to overwhelm agents with too many options.

> *"... as long as it's not too many buttons, like, probably less than five would be useful because that gives me more granular control versus having too many options of things to turn on or off... some of those aspects could be linked, like, probably you want x and y together and then a and b together, so being able to turn them both on and off at the same time would be sufficient."*

4.1.1 Design Iteration. We followed P0's suggestion to remove excessive information from the map. For example, in the airport navigation scenario, instead of showing annotation of each airline's check-in counter, we highlight the general check-in area during path planning and enable the text-reading feature to help agents find the correct desk. Note that relying on sign reading [99, 128] not only helps lower agents' cognitive load, but sometimes is also a necessity. For certain areas that are already challenging to navigate, such as security checkpoints, it is also often hard to find reliable maps.

Following P0's suggestion, the maximum number of buttons presented in each scenario was reduced from ten to five. We deleted the compass button and enabled compass on both Google maps and the live video feed permanently. Similarly, we altered the "change first-person views (side-by-side or overlay display)" button to an icon, which will be displayed on the top right of the camera feed when this feature is available on 3D maps. We removed the buttons enabling viewing the global map, destination, and current location on 3D maps in the office building scenario. Instead, we designed the function to double click the current location or destination to switch views on 3D maps, and zoom out to view global maps. Moreover, the buttons for recognition [59, 60, 83, 113] and staff recognition were combined.

4.2 Enhancing the Agent's Ability to Stay Ahead

One of the primary contributors to a quality experience for agents' clients that P0 identified is the agent "staying ahead" of PVI.

> *"... the agent is always supposed to be a few minutes ahead of the explorer."*

As mentioned in the previous subsection, P0 told us that she takes advantage of the first minute of the call to work ahead of PVI. She added that she also opportunistically prepares while PVI are waiting in queues.

> *"... then while they're in the queue, I might be doing the next step..."*

To further support the agent's ability to stay ahead, P0 recommended that we move our labeled distance bands (in Figures 2a-2b) further out than five to ten feet to allow for more agent processing time.

> *"I really liked the distance circles at the base of the video camera... however, the five and ten foot distance is actually– as an Aira agent, we're always trying to proactively see further because I need time to visually process information and spit it back out again... we try to project more to like 30-60 feet out, and then, because if I am noticing something 30-60 feet out, I have time to process and then tell the person that in 20 feet, for example, they will encounter an obstacle or whatever that might be."*

4.2.1 Design Iteration. To provide agents with more processing time, we modified the distance measure [57, 157] to detect and highlight obstacles at least thirty feet away, with intervals of ten feet.

While navigating, agents consult multiple sources of information and multitask continuously, splitting their attention [88]. Better understanding agents' desire to plan ahead and recognizing the challenge of keeping up with maps and the video feed simultaneously, we proposed a new way for agents to interact with 2D maps. This will enable agents to draw and highlight motion paths/plans on 2D maps manually. Then, CV will locate the PVI on the 2D maps by landmark recognition [59, 60, 83, 113] and scene text reading [146], and project walking directions and destinations from the map to the video feed. Thus, agents can plan a path early and then track the PVI on maps and reference the planned path in the live video feed rather than memorizing a path and trying to mentally contextualize it within the video.

4.3 Contextualizing Object Detection

One of RSA agents' top priorities is the monitoring and dictation of obstacle-related information [88]. In CV, this can be addressed by detecting objects in the PVI's path. However, P0 explained that agents are often not concerned with the details or classification of obstacles around PVI. Rather, she says that she and the PVI just want to know that an obstacle is nearby and where it is located.

> *"... the thing that is most important to me is that there's an obstacle there. Unless my explorer is actively looking for something, what that obstacle is matters less... The most important aspect is how far away am I from that obstacle and that there is an obstacle present..."*

She explains that providing unnecessary details about obstacles can overwhelm PVI's auditory channels, which are critical to their orientation and navigation skills.

> *"... to a person who's blind who's navigating... particularly in spaces that are very noisy... it can be a little disorienting if you're relying on your ears in order to travel through space, so knowing where those obstacles are allows me to provide that information without needing to tell you it's a cabinet..."*

On some occasions, however, identifying an obstacle can be important. P0 said that some objects are more likely to be relevant to a PVI's goals, so labels for some specific object types would be useful.

> *"... some types of obstacles are more important than others because the person is more likely to be looking to use that thing. So, for example, a trash can... I do have to have the person move the camera around looking for a tower shaped object, and sometimes I miss it, so then we have to go all the way back around again, so being able to identify maybe a hierarchy of objects would be very helpful versus having all objects available all the time."*

While annotating certain object types seems simple in principle, P0 says that which obstacles are of interest depends on the context and the task at hand. Sometimes priorities can be predictable, but agents get calls for a vast array of different activities.

> *"... we do many types of tasks in many different locations. Of course there are always patterns, so at the park, you're likely finding a restroom, finding a trash can, taking a walk through the park... but it could also be having a picnic with friends..."*

Because of the lack of certainty regarding what a PVI will be doing, P0 emphasized the importance of the ability to toggle features on and off at the agent's discretion.

4.3.1 Design Iteration. P0 mentioned that some objects are related to navigation, and that those should be labeled. Therefore, we developed a feature to detect relevant objects [59, 60, 83, 113], in which CV continuously detects objects in the camera feed and marks them on the map. Given that agents are interested in different objects (e.g., trash cans) at different times and that object detection algorithms may not be able to account for this variability, we also added a search bar at the top right of the map, where agents can enter the name of an object that they wish to locate. Then, CV will project the target object to the camera feed and highlight it so that agents can quickly find the desired object and guide the PVI to it.

Another related comment is that the importance of certain objects varies by context and task. We applied the detection of relevant objects to the walk in the park and grocery store scenarios, which can represent outdoor and indoor navigation respectively. When walking in a park, PVI are likely to look for a trash can, bench, or vending stand. In the grocery store scenario, agents can use this feature to find an aisle where desired items are located. If agents enter the item name in the search bar, the approximate location of the target item will be projected to the camera feed. This can be achieved by identifying objects with CV and matching annotations on 3D maps. Landmark recognition and scene text reading are also available in this scenario to facilitate easier interpretation of information on the signs.

4.4 Emphasizing the PVI's Video Feed

P0 also conveyed the importance of the live video feed to agents' work. While maps, satellite images, and other web-based information are useful, the video and audio feeds from the PVI's mobile device are the only sources of information that are real-time and guaranteed to be accurate and up to date.

> *"... the video feed itself is essentially my lifeline of information to what is happening in the PVI's location because, as a remote sighted agent, it is the most real-time information that I have..."*

Agents also pay close attention to PVI's video feeds because the feed gives them a first-person view of a PVI's surroundings, which allows them to make decisions based on their natural visual intuition. For example, P0 said that, at malls, she guides PVI to *"where I would expect an entrance to be"* based on the same visual cues that she would use to make sense of the location if she were there herself. Because of agents' dependence on the live video feed, P0 became concerned about the design ideas that involved overlays of AR features or other information on the video, as shown in Figures 1c, 2a, and 2b.

> *"So, in some of the scenarios, I noticed that there is a lot of information that has been superimposed on top of the video feed, which would make it, though very feature rich, difficult for me to actually see the location where the person is moving through space."*

She was wary of the overlaid information obscuring the video and making it more difficult for her to see the PVI's surroundings, possibly even such that it could threaten the safety of the PVI. She suggested separating some of the visual features from the video feed.

> *"The obstacle detection I think might be one of the aspects that might get very visually cluttered very quickly, so I don't know if it might be possible to maybe move obstacle detection out of the immediate video feed and onto the side piece?"*

The ability to toggle any features that may interfere with or produce an overlay on the video feed affects her position on their viability. As one example, if she were navigating a PVI through a parking lot with moving vehicles, she may

toggle AR directions and obstacle detection off so that she always has a clear view of the video feed and any moving cars.

4.4.1 Design Iteration. Because of agents' dependence on the live video feed, it is critical to minimize clutter on the video resulting from AR features or other information. Following this principle, we revised our design to pay special attention to the context and relevance of objects in navigation tasks, as well as the agents' needs.

For example, as explained in the previous section, CV is used to continuously detect relevant objects in the video. Instead of displaying all object labels on the video, we proposed to store the information in the map. We further provide a search bar for the agents to enter the name of a target object, and only project the target object to the camera feed.

As another example, instead of detecting and marking all moving objects (e.g., pedestrians, bikers, etc.) in the video, we focus on identifying people who are more likely to collide with the PVI [53]. This not only helps remove redundant information from the video, but is also a useful feature for ensuring PVI's safety.

4.5 Managing Risk in Navigation

P0 brought up PVI safety as a critical consideration in the development of new CV mediated RSA features. In the parking lot scenario, this meant identifying safe walking paths including sidewalks and crosswalks and path planning so as to minimize time spent outside of these designated areas.

> *"... best practice is always keep an explorer on a pedestrian walkway, so that would be a sidewalk, but that also means identifying in a parking lot where those crosswalks are..."*

In our prototype, we depicted CV recognizing and marking every pedestrian in the PVI's video feed. P0 pointed out that some pedestrians are more dangerous to the PVI's safety than others and that the relative risk can be determined visually. She said that people who see a PVI generally make an effort to move out of the way, but people that do not see the PVI because they are turned away, looking at their cell phone, or otherwise not paying attention do not.

> *"... I don't know if it would be possible to identify maybe, like, a distraction risk, but to me as an agent, those really are the individuals who are a higher threat versus somebody who is walking directly at the blind or visually impaired explorer."*

4.5.1 Design Iteration. To address these safety concerns, it is important to understand risk factors and safest practices. For example, P0 pointed out that people who do not see the PVI are more dangerous to the PVI's safety. Thus, we proposed to detect distracted people and alert the agent if a potential collision is imminent. For example, CV may detect pedestrians who are facing away from the PVI or taking photos in the revised walk in the park scenario.

4.6 Managing External Factors

Many other factors that are out of the agents' direct control affect the PVI experience of RSA and the possibilities for CV in RSA. For example, when presented with the rideshare scenario, P0 described how ridesharing services have been progressively restricting third party app integration capabilities. She explained that Aira has had some integration with popular ridesharing services but that functionalities have been becoming more limited and less useful.

> *"... both services have removed their sandbox [test environment], so engineers can't even play around with the app integration before having to deploy it, and also, they have started to block a lot of the incoming information."*

Agents can also no longer summon rides for PVI. For these reasons, P0 suggested that we abandon this scenario. We found her expert opinion on this issue to be compelling and agreed that the trend toward increasing integration restrictions presented a significant obstacle to the one design idea presented in the scenario, which relied on having the rideshare's location information. We then probed for whether there was another common and challenging scenario dissimilar to the others that we had presented. P0 recommended that we develop a grocery store scenario because of the pervasive lack of up-to-date store maps, the frequent movement of items and sections and difficulty of locating some items and sections, and the connectivity issues caused by large refrigerators. We believed that our design ideas would apply to a grocery shopping scenario in cogent ways and would manifest in unique manners compared to our other four scenarios. Additionally, we were motivated to analyze and include such a scenario because grocery shopping is a necessary and regular activity. We therefore substituted a grocery store scenario for the rideshare scenario for the design review study.

As P0 mentioned when discussing challenges in grocery stores and while reflecting on other scenarios, another problematic external factor is map accuracy. When we proposed that agents make use of interactive airport maps, P0 informed us that she has found that, *"Unfortunately, a lot of airport maps are intentionally incorrect for security..."* In particular, she noted that security checkpoints are often marked incorrectly. For that reason, she suggested that we not assume that all detailed information on airport maps is accurate and that we label only general areas.

Where possible, some agents do go out of their way to work around external constraints. For example, P0 told us that pet relief areas do not have standardized indicators and are rarely marked on maps. To provide a quality experience to PVI despite this challenge, P0 said that agents have begun keeping track of the locations of pet relief areas in airports by identifying them on maps and then storing those annotated maps so that other agents can reference them. They make this extra effort because *"it is so impacting for our customers."* CV may be able to help agents further overcome challenging external factors, for example by recognizing a variety of pet relief indicators.

4.6.1 Design Iteration. We replaced the rideshare scenario with a grocery store scenario. P0 indicated three major challenges in this scenario, namely the lack of the updated store maps, difficulty locating specific items, and the connectivity issues exacerbated by large refrigerators and metal displays. Similar to the office building scenario, we construct a 3D map of the store to address the lack of indoor maps in indoor navigation. Because the locations of individual items change frequently, the 3D maps of grocery stores will not include individual items but keep 3D structures of sections and aisles only.

The 3D map is also effective for localizing PVI under unstable network connections because the PVI's device only needs to send AR points, rather than a video feed. CV can localize the PVI by matching AR points with points on 3D maps. To further visualize the PVI's surroundings, the live video feed on the agent's dashboard can be augmented by the first-person view of a 3D map. As shown in Table 2, we also apply object detection, landmark recognition, and scene text reading to this scenario.

5 STUDY 2: DESIGN REVIEW STUDY

Next, we present our findings from the interviews with 12 professional agents and discuss the desirability and feasibility of our revised design ideas as well as agents' suggestions. The method of this study is described in Section 3.2.2.

Our bottom-up data analysis identified 15 themes in the findings of this study, which are organized into five high-level themes. These high-level themes solidify the first five themes in Section 4: *reducing the agent's cognitive load*, *enhancing*

the agent's ability to stay ahead, contextualizing object detection, emphasizing the PVI's video feed, and *managing risk in navigation*. Moreover, findings in this section supplement important context and details within each of these themes.

5.1 Reducing the Agent's Cognitive Load

Agents are tasked with processing a large volume of information while navigating PVI, including checking landmarks on the live video feed, referring to the map to localize and orient the PVI, describing surroundings, and guiding the PVI to avoid obstacles. Thus, it is vital to reduce agent cognitive load by providing concise, timely information.

5.1.1 Orientation. Seven agents (P2, P3, P5, P8, P9, P11, P12) indicated that orientation is one of the most difficult tasks in navigation. The compass allows agents to determine PVI's orientation in real time, saving them time and energy.

> *"Being able to orient and explore with the compass would make things far easier."* (P5)

Agents explained that orienting a PVI is generally their first task when they get a navigation call. Using their current tools, they cannot easily deduce a PVI's orientation, especially in homogeneous surroundings, like in a parking lot or at an intersection.

> *"... just orientation would be a huge help in parking lots, you know, because you always want to get them out of the parking lot as soon as possible... the first thing you want to do, like, know, is which way they're facing, and which way they're going..."* (P11)

Additionally, by projecting walking AR directions onto the video feed, there is no need to reorient the PVI if the call is disconnected or the PVI turns around.

> *"I think that's incredible because sometimes, if a call gets cut off, or if [a PVI] gets turned around, we have to orient them again, but if that line is there, all we have to do is turn around and be like, 'There it is.' So cool. I like this."* (P5)

5.1.2 Reducing Reliance on Scanning. Sometimes agents need PVI to scan an area with their camera in order to find objects or visual cues of interest. P6 gave an example of how he uses this strategy in grocery stores.

> *"First thing that I try and do is I get them as kind of far away from the aisle signs as I can and then angle their cameras up and then over so I can try and read each sign, each aisle marking, and, you know, that usually takes about five minutes if I'm lucky to get it done that fast."* (P6)

Guiding PVI to scan the appropriate areas at the correct angle can be challenging and time consuming. Agents believed that having the text read for them quickly and regardless of orientation would make navigation more efficient for both the agent and PVI, as it reduces reliance on precise scanning of the PVI's surroundings. Eight agents (P1, P2, P6, P7, P8, P10, P11, P12) considered it to be a beneficial feature.

> *"I love the signage because that is... another huge thing that a lot of times is hard for agents to guide a [PVI] to get their cameras to the right angle to get that picture and read those signs, so I love that."* (P8)

It can also be difficult for agents to identify information if the feed is fuzzy while the PVI moves the camera or if the connection is unstable.

> *"I love the idea of reading some of those signs because, again, sometimes the signal can be a little problematic, and so we have to do a lot of maneuvering with the cameras just to see those signs over each [aisle], and if we're reading it for the agent, I can save a huge amount of time."* (P1)

Two agents also suggested that the scene text reading feature could be utilized to smooth out the navigation process under an unstable network connection if the text could be read even when the video is of poor quality.

> *"If that video is going in and out, having it still pop up with information... would be very helpful as well, rather than, 'Okay, pause, let's wait for it to reload. Okay let's go again,' and so that would just speed up the whole process again."* (P3)

The first-person view constructed from the video feed embedded in a 3D map can also reduce the need for PVI to scan their surroundings by displaying out of frame objects.

> *"... then they'll be able to see things, potentially, that were outside of the camera view, which would help a lot... There's no guarantee that [PVI scanning is] going to give you the information that you need... so not having to rely so much on that, I think would be really great and really beneficial and make it a lot easier for the agent to know... what's in that direction."* (P8)

Furthermore, P10 described the view provided by the 3D map as more natural, as it is wider than the video feed alone and includes more information, closer to the way a sighted person would see their environment.

> *"... expanding the view... as if we're standing there because often... our vision is limited to this little section here... Sometimes the door that they need is just out of the view."* (P10)

5.1.3 Translation and Conversion. Aira serves clients in the US, the UK, Canada, New Zealand, and Australia, and within those countries, clients can be accustomed to different regions and use regional language. Some agents experienced cases in which they were unfamiliar with language or terms used, such as "level crossing," meaning an intersection with railroad tracks in Australia. Thus, three agents (P1, P6, P9) suggested adding real-time translation and a key of terms for the same objects in different cultures. P9 said that lack of familiarity with local terminology could be especially problematic when agents try to search for certain objects or landmarks.

> *"... if we're working with [PVI] who are in different countries or different cultures or something like that, their term for trash can... might be different... having to think of the word that we would use to search for that... that might take a little bit of extra time to find in that specific scenario..."* (P9)

Similarly, American agents struggle with converting distance from imperial to metric units, and P10 suggested that distances be able to be displayed in metric or imperial.

> *"... feet are easy for me as an American, but when we're working in Canada or countries that use the metric system, it's a little bit harder for my brain to switch as easily to those distances, so if those bands were shown in the correct measurements for them, that would make my life easier, for sure."* (P10)

5.2 Enhancing the Agent's Ability to Stay Ahead

Staying ahead is one factor that contributes to a seamless experience for PVI because, as one agent put it, *"some people just don't like waiting."* Agents said that distance estimation, path planning on 2D maps and 3D maps, and AR directions would help them to stay ahead of PVI.

5.2.1 Distance Estimation. It is difficult for agents to estimate distance through a video feed, especially considering differences in camera height and angle. We presented distance information in three different ways: as circular bands on the video feed radiating out from the PVI's position, as labels on obstacles and pedestrians, and as a grid overlaid

on maps. Nine agents (P1, P2, P5, P7, P8, P9, P10, P11, P12) gave positive feedback about the distance measurement features, and there were positive reactions to all three modes of distance information.

> *"That's... very helpful. I mean, distance measures, because that's, it's always tough, you know, and especially when you're trying to project things ahead of time."* (P2)

Being aware of the distances from the PVI to other things in the video feed helps agents to work ahead of PVI and prepare them for approaching obstacles, making them more confident in their navigation.

> *"... providing them, even providing with rough estimates... would again be a leap forward because it would give them a lot more confidence."* (P1)

Finally, calculating distances for agents frees them to focus on other aspects of their task and work faster.

> *"... the less that [agents] have to think about it, so that they can put that mental energy into other parts of their agent-ing, the better... the boxes that come up and tell them exactly how far it is, like, they don't have to think about it. They don't have to do any math. I don't have to look at any grids, you know, it's really straightforward."* (P8)

5.2.2 Path Planning and 3D Maps. Agents can employ path planning on both 2D maps and 3D maps. We received more positive feedback about 3D maps because they allow agents to path plan and explore the PVI's location when 2D maps are unavailable or the network connection is unstable (P1, P3, P5, P6, P10, P12).

> *"This is definitely a huge pain point as far as navigating in an unfamiliar building and kind of having to just take a look around, and just explore that with them, but if there was an option to have a kind of global map like you have here, already set up, that would be really great."* (P9)

Agents were particularly excited about the prospect of path planning using 3D maps in grocery stores. We learned from agents that just identifying the right aisles can take more than half of the total time that they spend assisting PVI in grocery stores.

> *"... it probably takes longer to find the aisle, and then that's probably the most work... with maps, they would be so much easier."* (P11)

P5 told us that grocery shopping can be greatly expedited if he is able to plan an efficient path through the store based on the items that PVI are looking for.

> *"A lot of times the [PVI] will send us that list of what they want to shop for. If I have the list... I can build my whole... route to get to wherever I have to get to right from the beginning."* (P5)

To further improve the agent's ability to stay ahead and reduce the PVI's waiting time, P3 suggested storing 3D maps in a database and automatically loading relevant maps based on the PVI's location.

> *"... if I try and search for a map online, I'm wasting their time, often, looking for the map that doesn't exist, so having a way to get to a 3D map quickly that's updated and accurate and helpful is good."* (P3)

Annotations on 3D maps make the path planning process more efficient because room numbers in office buildings and aisles and sections in grocery stores provide important details about building layouts and possible paths. Even beyond planning specific paths, agents said that learning about the general layout of locations like airports and grocery stores helps them to guide PVI more effectively.

> *"This map here, in conjunction with having the labels of the next slide, I think is very helpful just to get a sense for the location and the store... getting oriented with where the checkout area is, how the... aisles are formatted..."* (P9)

Agents considered crowdsourcing a promising way to construct and annotate 3D maps. This would entail agents and sighted volunteers scanning the insides of buildings and labeling relevant locations (e.g., service desks, bathrooms, pet relief areas, checkout counters). Based on their current practice of tracking pet relief areas in airports described by P0, we were not surprised to hear that agents would be willing to annotate maps.

> *"... a lot of [PVI] have guide dogs... It is very important for a lot of [PVI] to be able to get there [pet relief areas] and a lot of maps, they don't have that, so if we are able to, like, let's say that we encountered a relief station, that we can add a label in there and it will save it into the database..."* (P7)

In parking lot scenarios, planning a path to a specific entrance and navigating PVI in uniform surroundings can be challenging and time consuming for agents, as they may have to reference Google Maps, mall maps, and satellite views. Map alignment and identifying elements likely to be close to entrances using CV are potential solutions, especially for finding entrances that are unmarked on Google Maps and out of frame. Eight agents (P2, P3, P5, P8, P9, P10, P11, P12) reacted positively to the map alignment concept. In addition to conveniently finding entrances and planning paths, aligning maps provides agents with a comprehensive view of store locations in a consolidated fashion.

> *"Sometimes [agents] will have several maps open, because one map doesn't have all the information, so the ability to bring all that information into one place without them needing to have the two monitors and looking at three or four different maps would be amazing."* (P12)

When agents do have to find and reference multiple maps, it requires attention that could be better used to guide PVI and work ahead. Searching for and making sense of several different maps also demands a significant amount of time.

> *"Overlapping multiple maps to identify points of interest, such as entrances or specific stores, that's huge because that would save us a lot of time."* (P10)

In the same way that labeled 3D maps can familiarize agents with airports and grocery stores, agents also said that seeing these maps can help them get a feel for the layouts of malls.

5.2.3 Direction Tracking and Prediction. AR directions show walking directions using AR arrows projected onto the video feed and notify agents of direction changes in advance. With the help of AR directions, agents can deliver directional information farther ahead than they can manage on their own using the video feed. This feature is of value because agents prefer to dictate directions to PVI before they reach the points where they need to change directions so that they have extra time in case of connectivity-related lags.

> *"I do love the AR directions... sometimes with connectivity and things like that, you know, you have a little glitch, it's– we always try to give them information ahead of time."* (P2)

Having distances and directions provided to them also allows agents to direct PVI without disruption to their other tasks, such as paying attention to the video feed or describing surroundings.

> *"... it lets you quickly see where you need to go, which makes you focus more on the surroundings to give that person a more robust experience because then you're not wasting your time trying to figure out exactly where this turn is or anything like that."* (P4)

5.3 Contextualizing Object Detection

5.3.1 Selective Pedestrian, Object, and Obstacle Detection. Our design ideas included features for obstacle and pedestrian detection, object detection and recognition, landmark and signage recognition, scene text reading, and staff recognition. Agents expressed different views and preferences regarding these features. However, one thing is certain: there is no need to detect and identify everything in every scenario. *"Prioritizing and coming up with a few vital ones for each [scenario]"* is more sensible.

Agents were split on the pedestrian detection and tracking feature. P5 and P6 believed that it would help them to prevent collisions in highly trafficked areas.

> *"I think the obstacle detection is great in bigger cities and on walks, as you're saying, if there are crowds, everyone is on their phones, so not a lot of people are looking up, so you do have collisions..."* (P5)

Referring back to agent cognitive load, P6 told us that he would like to have pedestrians tracked for him so that he can focus on other mental tasks.

> *"It'd be incredibly helpful to have that, like open up that additional kind of processing power in the brain to not have to worry about [collisions]."* (P6)

Other agents were doubtful of the benefits of pedestrian detection and tracking. P7 was concerned that the feature could overwhelm her and overpopulate the video feed in busy locations.

> *"... if you have, like, 20 people... it'll be people, people, people, people, and then it'll, like, just cloud your view for navigation."* (P7)

P2 pointed out that most people will not be collision risks, so she tends not to worry too much about other pedestrians and their trajectories.

> *"A lot of times with pedestrians, especially when they're kind of walking towards each other, that's something that we tend not to describe because most of the time the pedestrian will see this person and move out of the way..."* (P2)

The variety of reactions and cases described with respect to pedestrian tracking suggest that it is a useful feature in some situations, but having the ability to quickly toggle it off is also important to agents.

Object detection and the search bar through which agents could have the CV highlight relevant objects in the feed were widely well-received. One common use case agents identified was locating trash cans for PVI with guide dogs who are walking in parks.

> *"The video signal itself that we're using might not be that good, and so with the computer to pick out something like the garbage can, where if the garbage can is dark green, it's blending in with the rest of the foliage, and so... the computer vision, that could help a lot."* (P1)

Agents also liked the idea that the search bar could help them to find specific objects through the visual "noise" in busy environments like crowded grocery store shelves or popular park paths.

> *"... since there's so much information that will pop up, especially like when reading those big signs at the store... if you are able to, like, quickly search and it'll show you right then and there, that's something that's very, very neat."* (P7)

The search bar, being a more passive feature (not active until an agent decides to utilize it), did not elicit objections like the other detection concepts.

Queue recognition again split the agents, with positive responses from six agents (P5, P7, P8, P10, P11, P12). Because of limitations on video feed quality and frame size, agents described struggling to identify queues from a distance, especially where the PVI needs to go to either walk around the crowd or join the end of the queue.

> *"The queue, like I said, I love that just because those are things that are standing still all the time, so... it's going to be easy for the camera recognition to just know, 'Okay, this is it. We need to avoid this area.'"* (P7)

P9, on the other hand, believed this feature to be unnecessary because she does not have a problem recognizing queues herself. Variance in reactions to queue recognition may stem from different experiences with PVI in locations that were more or less crowded and with signals that were more or less strong.

5.3.2 Scene Text Reading. Scene text reading was well liked in the grocery store scenario because of the connectivity problems common in grocery stores.

> *"... for the most part, we're able to have them look up and right and look at the sign, but it might not be clear, so if it were able to detect and read the sign even though the video feed's not clear, that would be excellent."* (P3)

P5 suggested that this feature could also be used to read "one-way" signage, which has been more common during the COVID-19 pandemic, in the airport scenario, where there are *"so many things happening at the same time..."* P5 expressed having difficulty examining surrounding details, especially when PVI are in a hurry and moving fast. He described an experience navigating a PVI through a one-way area in the wrong direction because he did not recognize one-way signage.

> *"... [PVI] are always always in a hurry... If you walk next door, which has happened before and has happened to agents, you walk and explore past a certain area, they are not allowed to walk back in that area. They have to walk all the way around, which is a huge, huge problem. It's very frustrating on both our parts and everything, but it's obviously more frustrating them."* (P5)

Although P5 pointed out the feasibility of recognizing one-way signage in the airport scenario, sign scene text reading was less appreciated for identifying signs that tend to be larger or follow a sequential numbering pattern. P8 suggested that airport gate numbers could be easily found online and therefore would not need to be read and displayed. P9 worried that the large number of signs in airports could lead to too much signage text being read and highlighted on the video feed, leading to excessive clutter.

> *"Having a lot of information at once kind of distracts from the central information that is most important."* (P9)

5.3.3 Staff Recognition. Staff recognition was also controversial but for different reasons. Five agents (P2, P3, P8, P10, P12) thought that the feature would be useful when PVI need assistance from staff members.

> *"[PVI] don't always need a staff member, but in the times that they do, it can be difficult to find those people, so if there was staff recognition, that would be, that would be pretty darn cool."* (P10)

Three other agents (P1, P5, P9) pointed out that finding staff is not a priority. P9 said that she just looks for a service desk when PVI need assistance. P1 and P5 explained that the feature could be useful but that many PVI expect agents to help them achieve their objective rather than find staff to help.

> *"One of the reasons they utilize our service is so they don't have to interact with a staff member for something like getting their luggage or anything like that... they consider that it means the agent's skill must be lacking if they have to find a staff member... It's still possible if they do definitely want assistance..."* (P1)

Two agents (P7, P8) expressed concern about staff recognition. They worried about CV's ability to correctly identify staff that may not be in full uniform and distinguish people in similar clothing from staff. P8 thought that agents may even become dependent on this feature and seek staff more often than necessary.

> *"One thing I think of with the staff recognition is I wouldn't want our agents to become dependent on that... I think that a lot of our agents would probably suggest that more often than is needed... I wouldn't want that to take away from the [PVI] feeling as independent as they want to..."* (P8)

5.3.4 Toggling Features On and Off. The variability in different agents' opinions on the features we presented and individual agents' preferences for various features in different scenarios highlight the importance of the ability to easily toggle features on and off.

> *"Again, just having options to be flexible with that. If that's something that automatically comes up and it's not really relevant or it's too distracting, that can take away from the work that is trying to be done..."* (P9)

P7 recommended saving agents' preferences regarding showing annotations, and P9 suggested *"having an option to to be able to focus in on one point"* and showing recognition in a selected area.

5.4 Emphasizing the PVI's Video Feed

Like P0, agents emphasized the centrality of the live video feed in the interviews. When online maps are unavailable or lack detail, agents' current method of compensating for missing information is to absorb as much information as possible from the video and match it to what they are able to see on maps. P4 provided an example of how he helps PVI navigate in newer, unmapped residential developments.

> *"I would say definitely relying more on the live video feed and then using the surrounding map of that property to lead them to either a driveway, sidewalk, or what have you, that we know does connect to that plot of land, and then it's strictly live video feed to kind of navigate through that new development..."* (P4)

Landmark recognition and scene text reading were well-liked because they facilitate easy recognition and interpretation of critical information that is missing from maps.

> *"I think that is a really, really great feature and that is one of the main tools that we do use to navigate when in an airport, if the airport map that we find online isn't very good or comprehensive..."* (P9)

Even if quality maps are available, some agents prefer to rely on the video feed to save time and effort associated with finding maps and localizing PVI on maps. P5 liked that scene text reading could replace maps by making directions easier to determine from the video feed.

> *"... I like to do as much as I can without the map because, as I said, it takes a little bit of time to bring it up and then to sit there and try to orient where they are and everything."* (P5)

P8 also had a positive response to scene text reading, especially the idea of *"all signage labels coming in the same format."* In most cases, formats and locations of signs are variable, which can cause agents to miss some signs if they are not in the color or spots that they expect. This feature helps agents to not miss information when signs look different or are *"hung on a wall or hung from the ceiling."*

Reiterating what P0 told us about the centrality of the video feed in navigation, P10 indicated that her main focus is on the real-time feed. For that reason, she liked that the distance bands were overlaid on the video so that she could see both distance and everything happening in the feed at the same time.

> *"... our main focus is on that camera feed because that's, that's real time, so the [distance] bands to me, that are actually on the camera feed would be most helpful... We try not to take our eyes off the camera..."* (P10)

5.5 Managing Risk in Navigation

PVI's safety is paramount to the agents' practice, especially during navigation. In parking lots, leading PVI to crosswalks or sidewalks is considered a best practice. We also received positive reviews of queue recognition, and some agents observed that it could be utilized for safety practices during COVID-19.

5.5.1 Parking Lots. To ensure the PVI's safety in parking lots, agents' current practice is to guide PVI in crosswalks, along the edges of parking lots, or on raised sidewalks between rows of parked cars, as stated by six agents (P2, P3, P8, P10, P11, P12).

> *"When we navigate explorers to parking lots, we try our very best to find sidewalks within the parking lot if we can, or if not, to kind of keep them along edges or the most direct walking route, whatever makes the most sense to keep them the safest."* (P10)

Agents wished to see CV support that likewise prioritized safe paths.

> *"Ideally, we always follow crosswalks... just for safety... but that's not necessarily the most direct route... so I would be hesitant to use something that didn't automate that way..."* (P3)

In our presentation, we depicted AR directions showing the most expedient path through a parking lot, which P3, P10, and others took issue with. P7 suggested adding the option to draw a path manually, as we presented with 2D maps in the airport navigation scenario.

> *"We're not supposed to, like, walk [PVI] through cars and stuff like that... if we have the option to, like, change it around just like we did in the... airport... that would be a great idea just because we still have to follow all of the rules that are, like, you know, being safe in a parking lot area."* (P7)

As in other scenarios, agents said that the video feed is of the utmost importance in parking lots for agents looking out for vehicles, people, and curbs. P3 and P10 therefore suggested making the AR directions semi-transparent in the parking lot scenario to prevent the overlays from obscuring any important visual information.

> *"I could see it being a distraction in a tense situation... If the overlay were transparent enough to not affect what's coming in from out of screen, because that's the most common thing in parking lots, is people pushing their carts, kids walking, cars moving..."* (P3)

5.5.2 Applications to the COVID-19 Pandemic. Two agents suggested extending the application of queue recognition to today's unusual circumstances. They believed that it could be useful for facilitating social distancing and preventing PVI from running into crowds during the pandemic.

> *"... with the, the social distancing and wanting, people wanting to stay, you know, far enough away from everybody... I think it would be useful."* (P8)

Moreover, P11 pointed out that she alerts PVI if *"there's a person not wearing a mask on the left in 15 feet."* Others suggested that distance bands set six feet out could help agents keep PVI from coming within six feet of others. P12

believed that it would be helpful to track and predict pedestrians' movement continuously during COVID-19 to avoid close contact.

> *"I think during social distancing, that [continuous movement tracking and prediction] would also be very helpful to know front, back, left, and right where they're at."* (P12)

Two agents (P5, P7) indicated that it is difficult to find open entrances during the pandemic because fewer entrances/exits are in operation. They valued the feature of finding invisible entrances and believed it could potentially alleviate challenges associated with new business operations during the COVID-19 pandemic.

> *"Especially when nowadays, I need to find more entrances because a lot of stores, because of COVID-19, have closed many of their entrances. They only have a particular entrance, I need that entrance."* (P5)

5.5.3 Unstable Network Connection. We proposed that agents may utilize a first-person view of 3D maps that keeps up with PVI's movement by matching AR points when PVI's video feeds are compromised by a poor network connection. Most agents liked having the option of using AR points to see where PVI are in buildings or stores, but some were concerned about not being able to see unpredictable obstacles from the 3D maps alone.

> *"... replacing the live feed with this video or this 3D map, it would take out the agent's ability to identify obstacles, so the [PVI] might run into things without the agent even knowing that they're there..."* (P8)

P9 informed us that agents are typically able to send PVI text messages even when the video cuts out, though. She figured that as long as she was able to tell PVI that she was no longer using the video feed and would not be able to tell them about unmapped obstacles that tracking them via AR points would be beneficial.

5.5.4 Further Opportunities. Agents also mentioned other challenging situations that our design ideas did not address that may pose risks to PVI. P4 and P12 indicated that elevation changes like small steps and ramps are difficult to distinguish in the video feed. P7 suggested detecting weather-related obstacles, such as mud, water, or snow on sidewalks. P9 said that agents might miss short or low objects near the ground depending on how PVI position their camera. If PVI hold the camera higher, *"we aren't able to see, like, if someone has a box kind of outside of their front door."* P3 and P9 recommended detecting the stanchion lines used to structure queues in airports since they are difficult to see in the video. PVI may run into these barriers because *"the cane goes underneath it, so they're thinking that they are free to continue traveling."*

6 DISCUSSION

In study 1, we presented low-fidelity prototypes for five real-world navigation scenarios to a RSA domain expert. Through a thorough analysis of the data, we developed six high-level themes that revealed how our design ideas might affect RSA practice and how they could be improved: (1) reducing the agent's cognitive load, (2) enhancing the agent's ability to stay ahead, (3) contextualizing object detection, (4) emphasizing the PVI's video feed, (5) managing risk in navigation, and (6) managing external factors.

After a design iteration based on these results, we evaluated the revised prototypes with 12 RSA agents in study 2. Using a bottom-up approach, we identified 15 themes, which can be organized into five high-level themes that reflect the first five found in study 1. Findings from study 2 supplemented details within each of these themes regarding how design ideas could smooth out the navigation process, how agents' views and preferences on some design ideas conflict, and additional challenging situations that our design ideas did not address.

In this section, we identify several opportunities for and limitations of CV mediated remote sighted assistance. The design concepts presented can be further improved through integration with PVI clients' profiles. We also discuss ways in which our findings can inform the development of CV applications for RSA beyond our design concepts, the ethics of a CV mediated RSA system, practical realities of crowdsourced map construction, the promising applications of the proposed designs to COVID-19 safety practices, the advantages of CV supporting RSA agents rather than directly supporting PVI, the effectiveness and challenges of using low-fidelity CV prototypes to present design ideas, and reasons for studying emerging CV technologies. Finally, we present the limitations of our research and the directions for future work.

6.1 Integrating Functions with PVI's Profiles

Each of Aira's visually impaired clients has a profile that agents can access containing relevant personal information (e.g., whether they are a cane or guide dog user) and RSA preferences (e.g., preference for more or less scenery description while navigating). If our proposed design ideas were integrated with these profiles, agents could employ CV features based on PVI's specific wants and needs to provide rich, meaningful experiences without having them take the time to fine tune the system for each PVI during every call. For example, as P0 suggests in Section 4.1, map annotations could include locations and labels that are personal to PVI, such as "my office." Preferred paths for frequent trips including points of interest along the way like restrooms, or pet relief areas for guide dog users, could also be saved to these profiles. Certain features could automatically be toggled on or off at the beginning of a call depending on the preferences recorded in the PVI's profile to quickly calibrate the system for the agent. For example, if a PVI's profile shows that they prefer not to consult staff for help, the system could toggle staff recognition off at the start of the call.

6.2 Opportunities for CV Algorithms in RSA

The high-level themes identified in the first and second studies provide general insights into the practices, goals, and challenges of remote sighted assistants and can therefore inform directions for CV support in RSA beyond the scope of the limited set of scenarios and design ideas presented here. Through the expert review and design review study, we learned that, in almost all cases, RSA agents wish to minimize their cognitive load while also continually working ahead of PVI. These themes, articulated by end users, define broad, relevant objectives for future CV technology in RSA. Further design and development of CV for RSA can be motivated by these objectives, and the potential value of early concepts can be assessed based on to what extent they advance these objectives.

First, tasks identified by agents as being cumbersome or cognitively demanding in Lee et al. [88] and the studies here, of which there are many, outline some specific opportunities for CV to enhance RSA services. For example, path planning with limited information regarding PVI's surroundings is a challenging and tedious problem. Currently, agents typically gather environmental information by asking PVI to scan their surroundings repeatedly. In the future, we envision that CV algorithms could learn to perform path planning with incomplete environmental information, leveraging a large set of historical video data associated with the actual paths taken by PVI. Researchers can also ideate and develop other CV functions that reduce agents' cognitive load and help them to work ahead of PVI and feel confident that such systems will be of some value, even if the jobs that they do are not yet documented agent challenges.

Second, the fact that the need to identify various types of objects is contextual represents further considerations, and some opportunities, for CV augmentations of RSA technologies. Beginning with object detection, which objects are of interest depends heavily on the scenario, context, and the preferences of the PVI and agent, and labeling all visible objects can result in information overload. However, given the large number of influential factors, it is virtually

impossible to pre-determine which objects to identify. While the association with PVI's profiles may help, it would be laborious to explicitly specify one's preference for a large range of possible activities. Thus, it is desirable to learn such information on-the-fly as we deploy systems in real-world applications. To this end, AI can be employed to automatically adjust the importance of features for each agent and PVI. Further, AI can be used to mine frequent patterns in user data and recommend additional objects to discover, similar to systems that recommend products based on a consumer's shopping history.

Likewise, the importance of the raw video feed to agents' practice suggests some constraints and opportunities for CV in RSA. Agents were critical of features that obscured or cluttered the video feed in dynamic (navigational) scenarios despite their potential to consolidate relevant information. CV designers should therefore explore methods of presenting information that do not interfere with the video feed but that still save agents time and do not add to their cognitive load (e.g., placing directions just above the video feed on the RSA interface). Researchers may also consider using CV technologies to enhance the user's video. For example, rather than adding opaque text tags below or next to signage, a system might read text from signage and portray the text clearly on top of the signage text in a similar color and style to clarify the text in a less distracting way.

Finally, future CV designs for RSA, especially applications that make decisions for agents (e.g., path planning functions), should more intentionally promote PVI's safety. For instance, per agents' feedback, path planning applications should prioritize safety over expedience by maximizing the use of marked crosswalks and sidewalks. Other technologies may be developed to identify or track safety hazards, such as puddles (as suggested by P7), large cracks in sidewalks, or other obstacles that are near or on the ground and leave the video frame as the user approaches them. Safety is of the utmost importance to agents (and presumably to RSA users as well), so CV applications that take some of the safety monitoring burden off of agents or extend their ability to protect PVI's safety is another promising design direction.

6.3 Collaborative Map Construction

We proposed constructing and annotating indoor 3D maps using crowdsourcing (Section 3.1) due to crowdsourcing's potential to make indoor 3D maps available and keep them up to date without significant financial investment from an RSA platform. Although prior work attests to the feasibility and importance of crowdsourcing for 3D map construction (Section 2.5), there are also some limitations of crowdsourcing 3D map construction. First, although a crowdsourcing system provides a convenient way to collect 3D map and label data, it fails to guarantee data quality. In terms of the quality of map data, one potential issue is loss of detail, such as a missing scan of an aisle in a grocery store. We envision that this problem might be gradually mitigated over time by the strength of crowdsourcing, merging information of the same location and supplementing more details. However, label data requires additional verification. Our system relies solely on manual labeling, where errors and imperfections are inevitable. Inaccurate labels influence the functionality of other features included in our design ideas. For example, spelling errors in labels and incompleteness of label data can affect the performance of a search bar for looking up relevant objects. Inappropriate placement of labels might occlude other important information on the 3D maps. Previous studies indicate that intrinsically motivated crowd workers give extraordinary effort in terms of both quantity and quality, whereas extrinsically motivated crowd workers that value monetary reward will prioritize completion speed instead of quality [118, 135]. Researchers hypothesized that increasing the intrinsic motivation of a task (e.g., altruism) may succeed in improving quality [118]. Sometimes even trustworthy crowd workers may provide inaccurate responses because of their inattentiveness, fatigue, or boredom [52], or because of poor instructions or task design [118]. The verification of accuracy can be supported by mechanisms studied in the area of CV, specifically obtaining high-quality labels for training object detection algorithms [136, 149]. To avoid errors

attributed to the poor instructions, we can provide detailed instructions of what to label (e.g., contextualization of object detection) and how to label (e.g., drawing bounding boxes around objects [136]).

Second, data age is an additional limitation. Unlike other crowdsourcing projects based on officially maintained datasets (e.g., Google Street View) [65, 66, 121], this proposition depends on volunteers creating 3D maps and editing labels. As we learned from agents in this study, maps lose their usefulness when they are out of date, and some types of locations (e.g., grocery stores) change frequently. Maintenance of useful 3D maps may therefore be difficult to ensure and require continuous efforts by crowd workers. Similar to approaches to improving data quality, we can seek to maximize and emphasize intrinsic motivation, particularly framing the task of collecting map and label data as helping PVI and RSA agents. Based on prior research [118, 135], we believe that intrinsically motivated volunteers would be willing to update 3D map and label information once they find out it is out of date, thus alleviating this problem to some degree. Institutional motivations for businesses, organizations, and local governments may also apply to this issue.

6.4 Extending the Proposed Design to COVID-19 Safety Practices and Beyond

Our scene text reading and 3D map annotation concepts can be applied to COVID-19 safety practices. One-way walking paths are one of the protocols currently being used to minimize close contact. We can use scene text reading to recognize one-way signage and direction arrows on the ground, especially for indoor navigation (Section 5.3.2). Finding entrances/exits in operation is challenging during the COVID-19 pandemic. Apart from utilizing the feature locating invisible entrances (Section 5.5.2), crowdsourcing is a relevant promising direction to update entrance/exit information. Sighted volunteers (e.g., government entities, employees of businesses) can upload annotations for entrances/exits to 3D maps to expedite the identification of viable entrances and exits for PVI.

The proposed design ideas can also be implemented for social safety practices beyond current pandemic conditions. Agents suggested that queue recognition and distance bands could facilitate social distancing and prevent PVI from running into crowds (Section 5.5.2), which may be desirable for a variety of reasons (e.g., avoiding distracted pedestrians and collisions, preventing unwanted touching of guide dogs). These design ideas can also help PVI avoid shoulder surfers, a criminal practice involving invasion of personal privacy and theft of personal data (e.g., passwords, ATM PIN). This common crime can be committed easily by spying over a PVI's shoulder when they use a laptop, ATM, kiosk, or other electronic devices in public. Design ideas that help maintain social distancing can assist PVI with keeping distance from pedestrians, thereby preventing shoulder surfers.

6.5 The Shift from CV Assisting PVI to CV Assisting RSA Agents

Our discussions with agents supported our hypothesis regarding the increased viability of CV assisting RSA agents compared to CV directly assisting RSA users. Agents explained how they often use their judgement and make frequent adjustments based on their context and PVI's changing preferences and needs to deliver RSA. They are also able to quickly recognize when a map is not up to date or when computer-generated directions might pose a safety hazard and use their problem-solving skills to deal with such situations. Human cognition and judgement is central to agents' practice and what agents classify as a quality RSA experience.

Throughout the interviews, agents pointed out several nuances to their practice that demonstrate how agents filter available information to suit particular scenarios and PVI's goals. For example, P0 said that she is more wary of unnecessarily verbally identifying objects if the PVI's location is especially noisy, a characteristic that might be transient (Section 4.3). Additionally, some agents opt to not search for a map if they feel like they can navigate the PVI's location using the video feed alone to save time. The rapid decisions that agents make to optimize their assistance

based on their subjective interpretation of the entire context are surprisingly common. Any CV algorithm that accounts for as many factors as agents do in decision making would be extraordinarily complex. Agents can use the continuous stream of visual cues from the video feed, outside information (e.g., maps), their understanding of their abilities and what the current client needs, and their ideas about what certain PVI might be interested in based on their personalities to judge what information is useful and what is not on-the-fly. A CV system, on the other hand, may overwhelm PVI or create poor experiences because it is not able to adapt as quickly as human agents. CV supporting agents can enhance agents' information collection and interpretation abilities and speed while still delegating them decision making power. No agents suggested that any feature provide information directly to PVI, even though that could potentially reduce their own cognitive load. Agents seemed to recognize the limitations of CV and their would-be role as a safety net in an RSA system with CV, as they often reacted positively to CV features but emphasized the importance of being able to quickly toggle features on and off or ignore the system to effectively adapt and maintain their standard of quality, personalized assistance.

6.6 Effectiveness and Challenges of Low-Fidelity Prototypes

In this paper, we prototyped and presented CV designs as low-fidelity, static, nonfunctional representations of a RSA interface. We evaluated the desirability of the prototyped CV concepts and user interface before putting extensive efforts into implementing CV technologies in a high-fidelity prototype or even creating a wireframe. Because we conducted the design review with RSA experts, participants were familiar with the cases presented and the challenges targeted. They demonstrated a thorough understanding of the design ideas despite the prototypes being low-fidelity. This approach was therefore appropriate and turned out to be highly effective and efficient for our purposes of reviewing core functionalities rather than design details.

First, by working with low-fidelity prototypes, we were able to explore the potential of less mature technologies while exercising a high degree of control over their presentation. Viewing nonfunctional prototypes, agents were not distracted by technological flaws and provided focused and constructive feedback on the usefulness of the design concepts. Second, the low prototyping cost allowed us to evaluate bolder design ideas and iterate quickly. Third, the low-fidelity prototypes made agents feel comfortable critiquing them honestly, knowing that they were not asking us to rethink months of work, evidenced by P0's recommendation to get rid of one of the five scenarios presented to her.

However, there are also challenges to portraying CV in low-fidelity prototypes. First, finding appropriate scenes and pictures is difficult. The researchers need to ensure that the low-fidelity prototype looks realistic enough that participants can look past its limited interactivity and react constructively to it. To achieve this goal, the researchers searched for visual resources online and filtered them in terms of size, angle, resolution, brightness, and content. For example, we illustrated queue recognition by highlighting a group of people standing still in front of airport check-in desks. When searching for pictures of airport check-in areas, the researchers did not select pictures that were too bright or too dim, pictures without queues of passengers, or images that did not appear to be taken from between waist and eye level. Second, although recognizing and labeling items on pictures manually is necessary, it is laborious when targets are numerous, clustered, and hard to identify. It is also challenging to coordinate and illustrate several CV capabilities simultaneously (e.g., constructing an image depicting obstacle detection, pedestrian detection, distance measurement, and AR directions simultaneously). In this case, the researchers used various colors and shapes to demonstrate different CV capabilities. Colors and shapes were chosen to be distinctive so that participants could distinguish labels for different features from the background and from each other easily. In addition to selecting proper colors and shapes for annotations, it is difficult to create, edit, or delete a label if it overlaps with the other one.

6.7 Reasons for Studying Emerging Technologies

While some of the technologies underlying the CV functionalities designed and prototyped here are not yet mature and ready for deployment, we found it valuable to explore future possibilities for CV in RSA with RSA experts. By studying emerging technologies in this way, we learned about general objectives and considerations for any type of CV implementation in RSA (Section 6.2), and we have defined directions for future CV research and development that have commercial desirability. Understanding which emerging technologies are in demand from various sources can influence CV research priorities and give researchers an idea of in what kinds of applications and domains CV technologies may be used, which may inform their development.

6.8 Limitations and Future Work

We had the unique opportunity to work with some of the few people who have specialized RSA expertise; the people we interviewed are literally the prospective users of the services we prototyped. However, our participants were all professionals working within the same RSA platform. Thus, they might be homogeneous in ways we do not intend or even understand. RSA is still a start-up area, but as it develops, we need to be on the lookout for opportunities to access other groups of RSA professionals.

We tried to address diverse settings in the scenarios represented in our prototype, and agents regarded them as typical and affirmed that we had addressed *"huge pain points"* in their professional practice. Still, we examined a relatively small set of real-world navigation tasks. This kind of limitation is inevitable in early-stage research. Broadening the set of scenarios we address and the diversity of navigational interventions we envision and investigate is an important future trajectory.

It is a limitation of our work that we designed and studied low-fidelity prototypes; through accompanying narrative we described interaction scenarios, but our participants were not able to actually interact with the computational support suggested by the prototypes. Our future direction is to implement and study higher fidelity prototypes and systems to more thoroughly explore and develop the functionality of CV mediated remote sighted assistance with feedback from not only RSA agents but also PVI.

Prior work [18] has investigated the privacy concerns of PVI associated with using RSA, such as inadvertently revealing sensitive and personally identifiable information to RSA agents, including medical prescriptions, credit card numbers, and emails. These privacy concerns are valid and important but are beyond the scope of this particular work, which has focused on agent's perceptions of early-stage conceptual design ideas. We also believe that personally identifiable information is protected at least in part by the professional RSA organization, Aira, where our participants work. Aira has a privacy policy and trained RSA professionals, characteristics that engender "institutional trust," as proposed by Akter et al. [18]. Although trained RSA professionals generally garner more trust than untrained RSA volunteers, these privacy concerns exist and warrant greater consideration in future studies of professional RSA services.

We understand and acknowledge the limitations of current technology in the context of the proposed design ideas. The concepts we presented integrate mature and emerging CV technologies that we envision will support RSA agents in navigation tasks and provide a better user experience for PVI. However, leveraging so many capabilities simultaneously requires a great deal of processing power and time with current technology. It could slow down RSA systems and exacerbate latency during poor network connections, though processing would occur on the agent's side, not the PVI's. We expect advancement in hardware and computation platforms (e.g., 5G networks, cloud computing) to continue to

expand the boundaries of the possible and grow increasingly close to meeting the computational needs of the designs proposed here.

ACKNOWLEDGEMENTS

We thank Aira for providing us with access to their RSA agents and the agents for participating in the interview study. This research was supported by the US National Institutes of Heath, National Library of Medicine (R01 LM013330).

REFERENCES

[1] 2021. Aira. https://aira.io/.
[2] 2021. ARCore. https://developers.google.com/ar.
[3] 2021. Autour. http://autour.mcgill.ca/en/.
[4] 2021. Be My Eyes. https://www.bemyeyes.com/.
[5] 2021. Google Maps. https://maps.google.com.
[6] 2021. Google Maps - Transit & Food. https://apps.apple.com/us/app/google-maps-transit-food/id585027354.
[7] 2021. Indoor Positioning System, Indoor Mapping SDK | LocusMaps. https://locuslabs.com/locusmaps/.
[8] 2021. Introducing ARKit 4. https://developer.apple.com/augmented-reality/arkit/.
[9] 2021. John F. Kennedy International Airport Terminal 4 - LocusLabs. https://tinyurl.com/2mj5pyne.
[10] 2021. One Mile at a Time. https://onemileatatime.com/wingtips-lounge-jfk-review/.
[11] 2021. OpenStreetMap. https://www.openstreetmap.org/.
[12] 2021. The Points Guy. https://thepointsguy.com/news/klm-789-economy-jfk-ams/.
[13] 2021. The Seeing Eye GPSTM App for cell-enabled iOS devices. http://www.senderogroup.com/products/shopseeingeyegps.html.
[14] 2021. TapTapSee - Blind and Visually Impaired Assistive Technology. https://taptapseeapp.com/.
[15] 2021. UPI.com. https://tinyurl.com/37u3ksp8.
[16] 2021. Welcome to Google Maps Platform - Explore where real-world insights and immersive location experiences can take your business. https://cloud.google.com/maps-platform/.
[17] 2021. Welcome to OpenStreetMap! OpenStreetMap is a map of the world, created by people like you and free to use under an open license. https://www.openstreetmap.org/.
[18] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. 2020. " I am uncomfortable sharing what I can't see": Privacy Concerns of the Visually Impaired with Camera Based Assistive Applications. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 1929–1948.
[19] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–971.
[20] Moustafa Alzantot and Moustafa Youssef. 2012. Crowdinside: Automatic construction of indoor floorplans. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. 99–108.
[21] Mauro Avila, Katrin Wolf, Anke Brock, and Niels Henze. 2016. Remote assistance for blind users in daily life: A survey about Be My Eyes. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 1–2.
[22] Yicheng Bai, Wenyan Jia, Hong Zhang, Zhi-Hong Mao, and Mingui Sun. 2014. Landmark-based indoor positioning for visually impaired individuals. In *2014 12th International Conference on Signal Processing (ICSP)*. IEEE, 668–671.
[23] Lamberto Ballan, Francesco Castaldo, Alexandre Alahi, Francesco Palmieri, and Silvio Savarese. 2016. Knowledge transfer for scene-specific motion prediction. In *European Conference on Computer Vision*. Springer, 697–713.
[24] Przemyslaw Baranski and Pawel Strumillo. 2015. Field trials of a teleassistance system for the visually impaired. In *2015 8th International Conference on Human System Interaction (HSI)*. IEEE, 173–179.
[25] Michela Bertolotto, Gavin McArdle, and Bianca Schoen-Phelan. 2020. Volunteered and crowdsourced geographic information: the OpenStreetMap project. *Journal of Spatial Information Science* 2020, 20 (2020), 65–70.
[26] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 65–72.
[27] BlindSquare. 2020. BlindSquare iOS Application. https://www.blindsquare.com/.
[28] Francisco Bonin-Font, Alberto Ortiz, and Gabriel Oliver. 2008. Visual navigation for mobile robots: A survey. *Journal of Intelligent and Robotic Systems* 53, 3 (2008), 263.
[29] Erin Brady, Jeffrey P Bigham, et al. 2015. Crowdsourcing accessibility: Human-powered access technologies. *Foundations and Trends® in Human–Computer Interaction* 8, 4 (2015), 273–372.
[30] Erin L. Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual challenges in the everyday lives of blind people. In *2013 ACM SIGCHI Conference on Human Factors in Computing Systems*, Wendy E. Mackay, Stephen A. Brewster, and Susanne Bødker

(Eds.). ACM, 2117–2126.

[31] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge J. Belongie. 2010. Visual Recognition with Humans in the Loop. In *11th European Conference on Computer Vision*. 438–451.

[32] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[33] Nama R Budhathoki and Caroline Haythornthwaite. 2013. Motivation for open collaboration: Crowd and community models and the case of OpenStreetMap. *American Behavioral Scientist* 57, 5 (2013), 548–575.

[34] Andrius Budrionis, Darius Plikynas, Povilas Daniušis, and Audrius Indrulionis. 2020. Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review. *Assistive Technology* (2020), 1–17.

[35] M Bujacz, P Baranski, M Moranski, P Strumillo, and A Materka. 2008. Remote guidance for the blind—A proposed teleassistance system and navigation trials. In *2008 Conference on Human System Interactions*. IEEE, 888–892.

[36] Michele A Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: challenges and opportunities. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 135–142.

[37] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian D. Reid, and John J. Leonard. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robotics* 32, 6 (2016), 1309–1332.

[38] Victor Campmany, Sergio Silva, Antonio Espinosa, Juan Carlos Moure, David Vázquez, and Antonio M López. 2016. GPU-based pedestrian detection for autonomous driving. *Procedia Computer Science* 80 (2016), 2377–2381.

[39] John Carroll and Amy Aaronson. 1988. Learning by Doing with Simulated Intelligent Help. *Commun. ACM* 31, 9 (Sept. 1988), 1064–1079. https://doi.org/10.1145/48529.48531

[40] John M. Carroll, Sooyeon Lee, Madison Reddie, Jordan Beck, and Mary Beth Rosson. 2020. Human-Computer Synergies in Prosthetic Interactions. *IxD&A* 44 (2020), 29–52. http://www.mifav.uniroma2.it/inevent/events/idea2010/doc/44_2.pdf

[41] Corey D. Chandler, Gloria Lo, and Anoop K. Sinha. 2002. Multimodal Theater: Extending Low Fidelity Paper Prototyping to Multimodal Applications. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI EA '02)*. Association for Computing Machinery, New York, NY, USA, 874–875. https://doi.org/10.1145/506443.506642

[42] Babar Chaudary, Iikka Paajala, Eliud Keino, and Petri Pulli. 2017. Tele-guidance based navigation system for the visually impaired and blind persons. In *eHealth 360°*. Springer, 9–16.

[43] Si Chen, Muyuan Li, Kui Ren, and Chunming Qiao. 2015. Crowd map: Accurate reconstruction of indoor floor plans from crowdsourced sensor-rich videos. In *2015 IEEE 35th International conference on distributed computing systems*. IEEE, 1–10.

[44] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli. 2013. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences* 110, 30 (2013), 12186–12191. https://doi.org/10.1073/pnas.1221464110 arXiv:https://www.pnas.org/content/110/30/12186.full.pdf

[45] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*. 2366–2374.

[46] Wael Elloumi, Kamel Guissous, Aladine Chetouani, Raphaël Canals, Rémy Leconge, Bruno Emile, and Sylvie Treuillet. 2013. Indoor navigation assistance with a Smartphone camera based on vanishing points. In *International Conference on Indoor Positioning and Indoor Navigation*. IEEE, 1–9.

[47] Wafa Elmannai and Khaled M. Elleithy. 2017. Sensor-based assistive devices for visually-impaired people: Current status, challenges, and future directions. *Sensors (Basel, Switzerland)* 17 (2017).

[48] Nobuo Ezaki, Kimiyasu Kiyota, Bui Truong Minh, Marius Bulacu, and Lambert Schomaker. 2005. Improved text-detection methods for a camera-based text reading system for blind persons. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 257–261.

[49] Navid Fallah, Ilias Apostolopoulos, Kostas Bekris, and Eelke Folmer. 2012. The user as a sensor: navigating users with visual impairments in indoor spaces using tactile landmarks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 425–432.

[50] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2018. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks* 108 (2018), 466–478.

[51] Giovanni Fusco and James M Coughlan. 2020. Indoor localization for visually impaired travelers using computer vision on a smartphone. In *Proceedings of the 17th International Web for All Conference*. 1–11.

[52] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1631–1640.

[53] Tarak Gandhi and Mohan M Trivedi. 2006. Pedestrian collision avoidance systems: A survey of computer vision based recent studies. In *2006 IEEE Intelligent Transportation Systems Conference*. IEEE, 976–981.

[54] Aura Ganz, Siddhesh Rajan Gandhi, James Schafer, Tushar Singh, Elaine Puleo, Gary Mullett, and Carole Wilson. 2011. PERCEPT: Indoor navigation for the blind and visually impaired. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 856–859.

[55] Aura Ganz, James M Schafer, Yang Tao, Carole Wilson, and Meg Robertson. 2014. PERCEPT-II: Smartphone based indoor navigation system for the blind. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3662–3665.

[56] Vanja Garaj, Rommanee Jirawimut, Piotr Ptasinski, Franjo Cecelja, and Wamadeva Balachandran. 2003. A system for remote sighted guidance of visually impaired pedestrians. *British Journal of Visual Impairment* 21, 2 (2003), 55–63.

[57] Andreas Geiger, Frank Moosmann, Ömer Car, and Bernhard Schuster. 2012. Automatic camera and range sensor calibration using a single shot. In *2012 IEEE International Conference on Robotics and Automation*. IEEE, 3936–3943.

[58] David Geronimo, Antonio M Lopez, Angel D Sappa, and Thorsten Graf. 2009. Survey of pedestrian detection for advanced driver assistance systems. *IEEE transactions on pattern analysis and machine intelligence* 32, 7 (2009), 1239–1258.

[59] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[60] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[61] GPS.gov. [n.d.]. GPS Accuracy. https://www.gps.gov/systems/gps/performance/accuracy/.

[62] Kristen Grauman and Bastian Leibe. 2011. Visual object recognition. *Synthesis lectures on artificial intelligence and machine learning* 5, 2 (2011), 1–181.

[63] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 5 (2019), 93:1–93:42.

[64] Richard Guy and Khai Truong. 2012. CrossingGuard: exploring information content in navigation aids for visually impaired pedestrians. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 405–414.

[65] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H Ng, and Jon E Froehlich. 2015. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with google street view: An extended analysis. *ACM Transactions on Accessible Computing (TACCESS)* 6, 2 (2015), 1–23.

[66] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 631–640.

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[68] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. 2017. Deep direct regression for multi-oriented scene text detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 745–753.

[69] Tsubasa Hirakawa, Takayoshi Yamashita, Toru Tamaki, and Hironobu Fujiyoshi. 2018. Survey on vision-based path prediction. In *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 48–64.

[70] Nicole Holmes and Kelly Prentice. 2015. iPhone video link facetime as an orientation tool: remote O&M for people with vision impairment. *International Journal of Orientation & Mobility* 7, 1 (2015), 60–68.

[71] Siyu Huang, Xi Li, Zhongfei Zhang, Zhouzhou He, Fei Wu, Wei Liu, Jinhui Tang, and Yueting Zhuang. 2016. Deep learning driven visual path prediction from a single image. *IEEE Transactions on Image Processing* 25, 12 (2016), 5892–5904.

[72] Ziad Hunaiti, Vanja Garaj, and Wamadeva Balachandran. 2006. A remote vision guidance system for visually impaired pedestrians. *The Journal of Navigation* 59, 3 (2006), 497–504.

[73] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227* (2014).

[74] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading text in the wild with convolutional neural networks. *International journal of computer vision* 116, 1 (2016), 1–20.

[75] Rabia Jafri, Syed Abid Ali, Hamid R. Arabnia, and Shameem Fatima. 2014. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. *The Visual Computer* 30, 11 (2014), 1197–1222.

[76] Sang-Hack Jung, Mayank Bansal, Bogdan Calin Mihai Matei, Jayan Eledath, Harpreet Singh Sawhney, Rakesh Kumar, and Raia Hadsell. 2014. Method and apparatus for real-time pedestrian detection for urban driving. US Patent 8,861,842.

[77] Rie Kamikubo, Naoya Kato, Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2020. Support strategies for remote guides in assisting people with visual impairments for effective indoor navigation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[78] Shaun K Kane, Brian Frey, and Jacob O Wobbrock. 2013. Access lens: a gesture-based screen reader for real-world documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 347–350.

[79] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. 2012. Activity forecasting. In *European Conference on Computer Vision*. Springer, 201–214.

[80] Eunjeong Ko and Eun Yi Kim. 2017. A vision-based wayfinding system for visually impaired people using situation awareness and activity-based instructions. *Sensors* 17, 8 (2017), 1882.

[81] Julian Francisco Pieter Kooij, Nicolas Schneider, Fabian Flohr, and Dariu M Gavrila. 2014. Context-based pedestrian path prediction. In *European Conference on Computer Vision*. Springer, 618–633.

[82] Adarsh Kowdle, Yao-Jen Chang, Andrew C. Gallagher, and Tsuhan Chen. 2011. Active learning for piecewise planar 3D reconstruction. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*. 929–936.

[83] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[84] Aliasgar Kutiyanawala, Vladimir Kulyukin, and John Nicholson. 2011. Teleassistance in accessible shopping for the blind. In *Proceedings on the International Conference on Internet Computing (ICOMP)*. The Steering Committee of The World Congress in Computer Science, Computer . . . , 1.

[85] Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. 2011. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 23–32.

[86] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 336–345.

[87] Sooyeon Lee, Madison Reddie, and John M. Carroll. 2021. Designing for Independence for People with Visual Impairments. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–19.

[88] Sooyeon Lee, Madison Reddie, Chun-Hua Tsai, Jordan Beck, Mary Beth Rosson, and John M Carroll. 2020. The emerging professional practice of remote sighted assistance for people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[89] Gordon E Legge, Paul J Beckmann, Bosco S Tjan, Gary Havey, Kevin Kramer, David Rolkosky, Rachel Gage, Muzi Chen, Sravan Puchakayala, and Aravindhan Rangarajan. 2013. Indoor navigation by people with visual impairment using a digital sign system. *PloS one* 8, 10 (2013).

[90] Ki-Joune Li and Jiyeong Lee. 2010. Indoor spatial awareness initiative and standard for indoor spatial data. In *Proceedings of IROS 2010 Workshop on Standardization for Service Robot*, Vol. 18.

[91] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2020. Deep learning for generic object detection: A survey. *International journal of computer vision* 128, 2 (2020), 261–318.

[92] Xu Liu. 2008. A camera phone based currency reader for the visually impaired. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. 305–306.

[93] Yuliang Liu and Lianwen Jin. 2017. Deep matching prior network: Toward tighter multi-oriented text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1962–1969.

[94] Ming-Chih Lu, Wei-Yen Wang, and Chun-Yen Chu. 2006. Image-based distance and area measuring systems. *IEEE Sensors Journal* 6, 2 (2006), 495–503.

[95] Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. 2018. Multi-oriented scene text detection via corner localization and region segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7553–7563.

[96] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia* 20, 11 (2018), 3111–3122.

[97] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M Kitani. 2017. Forecasting interactive dynamics of pedestrians with fictitious play. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 774–782.

[98] Roberto Manduchi, Sri Kurniawan, and Homayoun Bagherinia. 2010. Blind guidance using mobile computer vision: A usability study. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. 241–242.

[99] Marwan A Mattar, Allen R Hanson, and Erik G Learned-Miller. 2005. Sign classification using local and meta-features. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 26–26.

[100] Troy McDaniel, Kanav Kahol, Daniel Villanueva, and Sethuraman Panchanathan. 2008. 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, HAS 2008*. Association for Computing Machinery, Inc. 2008 1st Ambi-Sys Workshop on Haptic User Interfaces in Ambient Media Systems, HAS 2008 ; Conference date: 11-02-2008 Through 14-02-2008.

[101] Microsoft. [n.d.]. Seeing AI - Talking camera app for those with a visual impairment. https://www.microsoft.com/en-us/ai/seeing-ai.

[102] M. Murata, D. Ahmetovic, D. Sato, H. Takagi, K. M. Kitani, and C. Asakawa. 2018. Smartphone-based indoor localization for blind navigation across building complexes. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–10.

[103] Brian J Nguyen, Yeji Kim, Kathryn Park, Allison J Chen, Scarlett Chen, Donald Van Fossan, and Daniel L Chao. 2018. Improvement in patient-reported quality of life outcomes in severely visually impaired individuals using the Aira assistive technology system. *Translational Vision Science & Technology* 7, 5 (2018), 30–30.

[104] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 3 (2001), 145–175.

[105] Wanli Ouyang, Hui Zhou, Hongsheng Li, Quanquan Li, Junjie Yan, and Xiaogang Wang. 2017. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence* 40, 8 (2017), 1874–1887.

[106] Nektarios Paisios, Alexander Rubinsteyn, and Lakshminarayanan Subramanian. 2012. Exchanging cash with no fear: A fast mobile money reader for the blind. In *Workshop on Frontiers in Accessibility for Pervasive Computing. ACM*.

[107] Rémi Parlouar, Florian Dramas, Marc MJ Macé, and Christophe Jouffrais. 2009. Assistive device for the blind based on object recognition: an application to identify currency bills. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 227–228.

[108] J Eduardo Pérez, Myriam Arrue, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. 2017. Assessment of semantic taxonomies for blind indoor navigation based on a shopping center use case. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*. 1–4.

[109] Helen Petrie, Valerie Johnson, Thomas Strothotte, Andreas Raab, Rainer Michel, Lars Reichert, and Axel Schalt. 1997. MoBIC: An aid to increase the independent mobility of blind travellers. *British Journal of Visual Impairment* 15, 2 (1997), 63–66.

[110] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[111] Paymon Rafian and Gordon E Legge. 2017. Remote sighted assistants for indoor location sensing of visually impaired pedestrians. *ACM Transactions on Applied Perception (TAP)* 14, 3 (2017), 19.

[112] Santiago Real and Alvaro Araujo. 2019. Navigation systems for the blind and visually impaired: Past work, challenges, and open problems. *Sensors (Basel, Switzerland)* 19, 15 (02 Aug 2019), 3404. https://doi.org/10.3390/s19153404 31382536[pmid].

[113] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.

[114] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.

[115] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144.

[116] Sebastião Rocha and Arminda Lopes. 2020. Navigation based application with augmented reality and accessibility. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3383004

[117] Ranga Rodrigo, Mehrnaz Zouqi, Zhenhe Chen, and Jagath Samarabandu. 2009. Robust and efficient feature tracking for indoor navigation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 3 (2009), 658–671.

[118] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[119] Mary Beth Rosson and John M Carroll. 2009. Scenario-based design. In *Human-computer interaction*. CRC Press, 161–180.

[120] Manaswi Saha, Alexander J Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 222–235.

[121] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. 2019. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[122] Daisuke Sato, Uran Oh, Kakuya Naito, Hironobu Takagi, Kris Kitani, and Chieko Asakawa. 2017. NavCog3: An evaluation of a smartphone-based blind indoor navigation assistant with semantic features in a large-scale environment. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 270–279.

[123] Jürgen Sauer, Katrin Seibel, and Bruno Rüttinger. 2010. The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics* 41, 1 (2010), 130 – 140. https://doi.org/10.1016/j.apergo.2009.06.003

[124] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. 2009. Make3D: Learning 3D scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 5 (2009), 824–840.

[125] Davide Scaramuzza, Ahad Harati, and Roland Siegwart. 2007. Extrinsic self calibration of a camera and a 3d laser range finder from natural scenes. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 4164–4169.

[126] Stefano Scheggi, A Talarico, and Domenico Prattichizzo. 2014. A remote guidance system for blind and visually impaired people via vibrotactile haptic feedback. In *22nd Mediterranean Conference on Control and Automation*. IEEE, 20–23.

[127] Nicolas Schneider and Dariu M Gavrila. 2013. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *German Conference on Pattern Recognition*. Springer, 174–183.

[128] Huiying Shen and James M Coughlan. 2012. Towards a real-time system for finding and reading signs for visually impaired users. In *International Conference on Computers for Handicapped Persons*. Springer, 41–47.

[129] Roy Shilkrot, Jochen Huber, Connie Liu, Pattie Maes, and Suranga Chandima Nanayakkara. 2014. FingerReader: a wearable device to support text reading on the go. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 2359–2364.

[130] Sudipta N. Sinha, Drew Steedly, Richard Szeliski, Maneesh Agrawala, and Marc Pollefeys. 2008. Interactive 3D architectural modeling from unordered photo collections. *ACM Trans. Graph.* 27, 5 (2008), 159.

[131] Arnold W. M. Smeulders, Dung Manh Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. 2014. Visual tracking: An experimental survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 7 (2014), 1442–1468.

[132] C. Snyder. 2003. *Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces*. Elsevier Science. https://books.google.com/books?id=YbzBWfTHorQC

[133] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. 2016. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4697–4705.

[134] Microsoft Soundscape. 2020. A map delivered in 3D sound. https://www.microsoft.com/en-us/research/product/soundscape/.

[135] Osamuyimen Stewart, David Lubensky, and Juan M Huerta. 2010. Crowdsourcing participation inequality: a SCOUT model for the enterprise domain. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. 30–33.

[136] Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[137] Ender Tekin and James M. Coughlan. 2010. A mobile phone application enabling visually impaired users to find and read product barcodes. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 290–295.

[138] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*. 1904–1912.

[139] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5079–5087.

[140] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 2016. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*. Springer, 56–72.

[141] Edwin Tjandranegara. 2005. Distance estimation algorithm for stereo pair images. *ECE Technical Reports* (2005), 64.

[142] Barbara Tversky. 1993. Cognitive maps, cognitive collages, and spatial mental models. In *Spatial Information Theory A Theoretical Basis for GIS*, Andrew U. Frank and Irene Campari (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 14–24.

[143] Prashant Verma, Kushal Agrawal, and V Sarasvathi. 2020. Indoor navigation using augmented reality. In *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations*. 58–63.

[144] Robert A. Virzi, Jeffrey L. Sokolov, and Demetrios Karis. 1996. Usability Problem Identification Using Both Low- and High-Fidelity Prototypes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, British Columbia, Canada) *(CHI '96)*. Association for Computing Machinery, New York, NY, USA, 236–243. https://doi.org/10.1145/238386.238516

[145] Jacob Walker, Abhinav Gupta, and Martial Hebert. 2014. Patch to the future: Unsupervised visual prediction. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 3302–3309.

[146] Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*. IEEE, 1457–1464.

[147] Wei-Yen Wang, Ming-Chih Lu, Hung Lin Kao, and Chun-Yen Chu. 2007. Nighttime vehicle distance measuring systems. *IEEE Transactions on Circuits and Systems II: Express Briefs* 54, 1 (2007), 81–85.

[148] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. 2018. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7774–7783.

[149] Peter Welinder and Pietro Perona. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 25–32.

[150] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. 2013. Inferring. In *Proceedings of the IEEE International Conference on Computer Vision*. 2224–2231.

[151] TW Yang, K Zhu, QQ Ruan, and JD Han. 2010. Moving target tracking and measurement with a binocular vision system. *International journal of computer applications in technology* 39, 1-3 (2010), 145–152.

[152] Cong Yao, Xiang Bai, Nong Sang, Xinyu Zhou, Shuchang Zhou, and Zhimin Cao. 2016. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002* (2016).

[153] Shuai Yi, Hongsheng Li, and Xiaogang Wang. 2016. Pedestrian behavior understanding and prediction with deep neural networks. In *European Conference on Computer Vision*. Springer, 263–279.

[154] Chris Yoon, Ryan Louie, Jeremy Ryan, MinhKhang Vu, Hyegi Bang, William Derksen, and Paul Ruvolo. 2019. Leveraging augmented reality to create apps for people with visual disabilities: A case study in indoor navigation. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 210–221.

[155] Chien Wen Yuan, Benjamin V Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M Carroll. 2019. Constructing a holistic view of shopping with people with visual impairment: a participatory design approach. *Universal Access in the Information Society* 18, 1 (2019), 127–140.

[156] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. 2016. Is faster R-CNN doing well for pedestrian detection?. In *European conference on computer vision*. Springer, 443–457.

[157] Qilong Zhang and Robert Pless. 2004. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, Vol. 3. IEEE, 2301–2306.

[158] Shanshan Zhang, Jian Yang, and Bernt Schiele. 2018. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6995–7003.

[159] Xin Zhang, Yee-Hong Yang, Zhiguang Han, Hui Wang, and Chao Gao. 2013. Object class detection: A survey. *ACM Computing Surveys (CSUR)* 46, 1 (2013), 1–53.

[160] Yu Zhong, Walter S Lasecki, Erin Brady, and Jeffrey P Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2353–2362.

[161] Baoding Zhou, Wei Ma, Qingquan Li, Naser El-Sheimy, Qingzhou Mao, You Li, Fuqiang Gu, Lian Huang, and Jiasong Zhu. 2021. Crowdsourcing-based indoor mapping using smartphones: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing* 177 (2021), 131–146.

[162] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5551–5560.

[163] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*. 3357–3364.

[164] P. A. Zientara, S. Lee, G. H. Smith, R. Brenner, L. Itti, M. B. Rosson, J. M. Carroll, K. M. Irick, and V. Narayanan. 2017. Third Eye: A shopping assistant for the visually impaired. *Computer* 50, 02 (feb 2017), 16–24. https://doi.org/10.1109/MC.2017.36

[165] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2019. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055* (2019).