Opportunities for Human-AI Collaboration in Remote Sighted Assistance

Sooyeon Lee*†
Pennsylvania State University
University Park, PA, USA
sul131@psu.edu

Rui Yu*
Pennsylvania State University
University Park, PA, USA
rzy54@psu.edu

Jingyi Xie
Pennsylvania State University
University Park, PA, USA
jzx5099@psu.edu

Syed Masum Billah
Pennsylvania State University
University Park, PA, USA
sbillah@psu.edu

John M. Carroll
Pennsylvania State University
University Park, PA, USA
jmc56@psu.edu

ABSTRACT

Remote sighted assistance (RSA) has emerged as a conversational assistive technology for people with visual impairments (VI), where remote sighted agents provide realtime navigational assistance to users with visual impairments via video-chat-like communication. In this paper, we conducted a literature review and interviewed 12 RSA users to comprehensively understand technical and navigational challenges in RSA for both the agents and users. Technical challenges are organized into four categories: agents' difficulties in orienting and localizing the users; acquiring the users' surroundings and detecting obstacles; delivering information and understanding user-specific situations; and coping with a poor network connection. Navigational challenges are presented in 15 real-world scenarios (8 outdoor, 7 indoor) for the users. Prior work indicates that computer vision (CV) technologies, especially interactive 3D maps and realtime localization, can address a subset of these challenges. However, we argue that addressing the full spectrum of these challenges warrants new development in Human-CV collaboration, which we formalize as five emerging problems: making object recognition and obstacle avoidance algorithms blind-aware; localizing users under poor networks; recognizing digital content on LCD screens; recognizing texts on irregular surfaces; and predicting the trajectory of out-of-frame pedestrians or objects. Addressing these problems can advance computer vision research and usher into the next generation of RSA service.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Collaborative interaction*; • **Computing methodologies** → *Computer vision problems.*

*Both authors contributed equally to this research.
†Also with Rochester Institute of Technology.

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03…$15.00
https://doi.org/10.1145/3490099.3511113

KEYWORDS

people with visual impairments, blind; remote sighted assistance, conversational assistance, RSA; computer vision, artificial intelligence; camera, navigation, smartphone, augmented reality, 3D maps

ACM Reference Format:
Sooyeon Lee, Rui Yu, Jingyi Xie, Syed Masum Billah, and John M. Carroll. 2022. Opportunities for Human-AI Collaboration in Remote Sighted Assistance. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3490099.3511113

1 INTRODUCTION

Remote sighted assistance (RSA) has emerged as a conversational assistive technology for people with visual impairments (PVI) [74]. In RSA paradigms, a *user* with vision impairment uses their smartphones to establishes a video connection with a remote sighted assistant, namely, an *RSA agent* or simply an *agent*, who then interprets the video feed coming from the user's smartphone camera, while conversing with the user to provide assistance as needed or requested. Recently, a number of RSA services came out of academia, e.g., VizWiz [19], BeSpecular [61], Crowdviz [60], as well as of industry, e.g., TapTapSee [109], BeMyEyes [17], Aira [8].

Historically, RSA services have been developed based on users' needs and feedback in multiple trials [16, 27, 70, 91, 104]. The communication in early RSA services was unidirectional, i.e., from agents to users, and the scope was narrow (e.g., agents describe objects in static images). As these services have matured over time and gained popularity, both agents and users adopted new technologies, such as smartphones, two-way audio/text conversation, realtime camera feed, GPS, and Google Maps. As a result, the current RSA services have broadened the scope in achieving complex tasks, such as agents assisting PVI in navigating airports and crossing noisy intersections without veering.

With the increased task complexity, researchers [65, 74] have identified that reliance on smartphones' camera feed can be a limiting factor for the agents, affecting their performance and mental workload, which can subsequently degrade the user experience of PVI. To elaborate, Kamikubo et al. [65] and Lee et al. [74], who studied RSA services, reported several challenges for agents, such as agents' lack of confidence due to unfamiliarity with PVI's physical surroundings, lack of indoor maps with fine details, inability

to track the PVI continuously on static maps, difficulty in estimating objects' distances in the camera feed, and describing relevant objects or obstacles in realtime. However, these challenges are derived from the agents' perspective, largely overlooking the user experience of people with vision impairments who are the *true* users of RSA services. As such, the findings in prior work are likely to be incomplete. This paper draws on prior work to holistically understand the technical and navigational challenges in RSA from both the agents' and the users' perspectives. More specifically, we aimed at understanding two research questions: *What makes remote sighted assistance challenging? When this assistance becomes challenging to use?*

To that end, we employed two methodologies. *First*, we conducted a literature review to identify technical challenges in RSA services, which are mostly derived from the agents' point of view. *Second*, we conducted an interview study with 12 visually impaired participants who use RSA services, in order to understand navigational challenges from their standpoint. Based on these two studies, we then constructed an exhaustive list of technical and navigational challenges to expand prior work and outline how these challenges occur in different real-world navigation scenarios.

We organized technical challenges into four broad categories: agents' difficulty orienting and localizing the users; acquiring the users' surroundings and detecting obstacles; delivering information and understanding user-specific situations; and coping with a poor network connection and external issues. Additionally, we produced a list of 15 real-world scenarios (8 outdoor, 7 indoor) that are challenging for PVI to navigate. A sampler of these scenarios include taking a walk around a familiar area (e.g., park, campus); calling a ride-share and going to the pick-up location; navigating through parking lots or construction sites; finding trash cans or vending machines; navigating malls, hotels, airports, train platforms; and finding an empty seat in theaters or an empty table in restaurants.

Because many of our identified challenges are well-researched in computer vision (CV) and AI literature, we investigated whether CV-based techniques can assist RSA agents in our prior work [120]. Our prior work suggests that having 3D maps of the users' surroundings and the ability to continuously localize users on the maps can benefit RSA agents in addressing a subset of those challenges. Complementary to our prior findings, a deeper analysis presented in this paper reveals that some challenges in RSA are too complex and dynamic to be addressed by CV-based automated approaches, thereby warranting new development in Human-CV collaboration. We formalize this prospective development as five emerging problems in Human-AI collaboration: (1) making object recognition and obstacle avoidance algorithms blind-aware during navigation; (2) localizing users under poor networks; (3) recognizing content on digital displays; (4) recognizing texts on irregular surfaces (e.g., curved); and (5) predicting the trajectory of out-of-frame pedestrians or objects. We believe our problem formulation will inspire computer vision and HCI researchers to find new solutions, which can usher into the next generation of RSA service.

2 BACKGROUND AND RELATED WORK

2.1 Navigational Aids for People with VI

Navigation is the ability to plan and execute a route to a desired destination. It is essential to have a spatial representation of users' surroundings (i.e., digital maps, cognitive maps [113], building layouts), direction information, and continuous update of their location in that representation (localization) [94]. Over the last 70 years, researchers proposed many prototypes to aid people with VI in both outdoor and indoor navigation. In this section, we only review a subset of such prototypes that are widely used and run on smartphones (for a chronological review, see Real and Araujo [95]).

Smartphone apps for outdoor navigation rely on GPS sensors for localization and commercial map services (e.g., Google Map, OpenStreet Map) for wayfinding. For example, BlindSquare [20], SeeingEyeGPS [5], Soundscape [107], and Autour [2]. These apps are feasible to navigate large distances for people with VI by providing spatial descriptions and turn-by-turn directions through audio. However, they are not reliable in last-few-meters [101] due to a wide margin of error in GPS accuracy (±5m [46]).

The weaker GPS signal strength indoor is also a barrier to indoor navigation. To overcome this limitations, researchers have fused available smartphone sensors as alternatives for indoor navigation, such as motion sensors, Bluetooth [103], Infrared [75], NFC [40], RFID [39], sonar [32], beacon [47] and camera. Lack of sufficiently detailed indoor map data is the other challenge [76, 99]. To mitigate this challenge, researchers have proposed to construct indoor maps by understanding the semantic features of the environment (for a complete list, see Elmannai and Elleithy [34]). Unfortunately, these solutions require additional deployment and maintenance effort to augment the physical environment [44], as well as a significant bootstrapping cost for setting up databases of floorplan [35] and structural landmarks [13, 90]. Some solutions also require the users to carry specialized devices (e.g., an IR tag reader [75]). For these reasons, no single indoor navigation system is widely deployed.

2.2 RSA Services for People with VI

RSA service is an emerging navigational aid for people with VI [29]. The implementation of various RSA services differs in three key areas. (i) The communication medium between users and remote sighted assistants. Earlier prototypes used audio [91], images [19, 70], one-way video using wearable digital cameras [27, 43], or webcams [27], whereas the recent ones are using two-way video chat using smartphones [8, 16, 17, 59]; (ii) The instruction form. RSA services are based on texts [72], synthetic speech [91], natural conversation [8, 16, 17], or vibrotactile feedback [30, 104]). (iii) Localization technique. For example, via GPS-sensor, crowdsourcing images or videos [19, 71, 94, 125], fusing sensors [94], or using CV as discussed in the next subsection.

Researchers have studied the crowdsourced and paid RSA services. For the crowdsourced RSA services (e.g., TapTapSee [109], BeMyEyes [17]), researchers concluded that they are feasible to tackle navigation challenges for people with VI [11, 21]. However, potential issues in crowdsourced RSA services include that (i) users trust too much on subjective information provided by crowdworkers, and (ii) crowdworkers are not available at times [28]. Compared with the crowdsourced RSA services, Nguyen et al. [88] and Lee at

al. [74] reported that assistants of paid RSA services (e.g., Aira [8]) are trained in communication terminology and etiquette, which means they do not provide subjective information. Furthermore, they are always available. In this paper, we assume Aira or a similar RSA service exists to demonstrate our design.

2.3 Use of CV in Navigation for People with VI

Budrionis et al. [26] reported that CV-based navigation apps on smartphones are a cost-effective solution. Researchers have proposed several CV-based positioning and navigation systems through recognizing landmark (e.g., storefronts [101]) or processing of tags (e.g., barcodes [75, 110], QR codes [33, 68], color markers [80] and RFID [81]). CV techniques have also been applied to obstacle avoidance [67, 93], which ensures users to move safely during the navigation without running into objects. However, Saha et al. [101] who studied the last-few-meters wayfinding challenge for people with VI, concluded that for a deployable level of accuracy, using CV techniques alone is not sufficient yet. Our goal in this project is to use CV to assist sighted assistants (e.g., RSA agents), rather than people with VI, who could be vulnerable to inaccuracies of CV systems.

Another line of work is to develop autonomous location-aware pedestrian navigation systems. These systems combine CV with specialized hardware (e.g., wearable CV device [78] and suitcase [48]), and support collision avoidance. While these systems have expanded opportunities to receive navigation and wayfinding information, their real-world adaptability is still questionable, as Banovic et al. [14] commented that navigation environments in real-world are dynamic and ever-changing.

Lately, researchers are exploring the feasibility of augmented reality (AR) toolkit in indoor navigation, which is built into modern smartphones (e.g., ARKit [3] in iOS devices, ARCore [1] in Android devices). Yoon et al. [123] demonstrated the potential of constructing indoor 3D maps using ARKit and localizing users with VI on 3D maps with acceptable accuracy. Troncoso Aldas et al. [112] proposed an ARKit-based mobile application to help people with VI recognize and localize objects. Researchers found that AR-based navigation systems have the advantage of (i) a widespread deployment [98], (ii) providing a better user experience than traditional 2D maps [114], and (iii) freeing users' hands without the need of pointing the camera towards an object or a sign for recognition [37].

More recently, we explored the opportunity of utilizing computer vision technologies to assist sighted agents instead of users with VI [120]. We designed several use scenarios and low-fidelity prototypes and had them reviewed with professional RSA agents. Our findings suggested that a CV-mediated RSA service can augment and extend the agents' vision in different dimensions, enabling them to see further spatially and predictably, as well as keeping them stay ahead of the users to manage possible risks. This paper complements those findings by identifying situations where leveraging CV alone is not feasible to assist sighted assistants.

2.4 Collaboration between Human and AI

Despite recent advancements of CV, automatic scene understanding from video streams and 3D reconstruction remain challenging [87]. Factors, such as motion blur, image resolution, noise, illuminations variations, scale, and orientation, impact the performance and accuracy of existing systems [64, 87]. To overcome these challenges, researchers have proposed interactive, hybrid approaches that involve human-AI collaboration [22]. One representative of the approach is the human-in-the-loop framework. Branson et al. [23] incorporated human responses to increase the visual recognition accuracy. Meanwhile, they found that CV reduced the amount of human effort required. Similarly, researchers developed interactive 3D modeling in which humans draw simple outlines [106] or scribbles [69] to guide the process. They increased the accuracy of 3D reconstructions while considerably reducing the human effort.

Collaborative 2D map construction and annotation is the other example of Human-AI collaboration, where AI integrates and verifies human inputs. Systems have been developed for collaborative outdoor map construction (e.g., OpenStreetMap [4]) and indoor one (e.g., CrowdInside [10], SAMS [92], and CrowdMap [31]). Researchers also probed the use of collaborative 2D map construction and annotation in supporting navigational tasks for people with VI. For example, improving public transit [53] and sidewalk [83, 102] accessibility, and providing rich information about intersection geometry [52]. Guy and Truong [52] indicated that collaborative annotations represent information requested by users with VI and compensate for information not available in current open databases.

Although prior work supports the technological feasibility of collaborative mapping and annotation, the motivation and incentives of volunteers have been a concern surrounding collaborative map construction. Budhathoki and Haythornthwaite indicated that volunteers can be motivated by intrinsic (e.g., self-efficacy and altruism) or extrinsic (e.g., monetary return and social relations) factors. In contrast, all volunteers are equally motivated in terms of a personal need for map data [25].

3 IDENTIFYING NAVIGATION CHALLENGES IN RSA: LITERATURE REVIEW

We aimed to understand the navigation challenges in RSA from two different perspectives, namely, the agents' and users' perspectives. This section presents a literature review that produces a list of such challenges from the agents' perspective.

We used Google Scholar (GS) to create an initial corpus containing papers from diverse sources. First, we defined a list of phrases specific to remote assistance for people with VI. More specifically, we considered these phrases: "visual impairment teleassistance", "visual impairment tele-guidance", "visual impairment remote guidance", "visual impairment O&M remote", "visual impairment O&M remote video", "visual impairment remote assistance", and "visual impairment remote sighted assistance".

We scraped top-10 search results returned by GS for each phrase – totaling 70 papers for all phrases. We restricted to top-10 results because we observed that (i) these results appeared on the first page of GS, indicating their high relevance, and (ii) results on subsequent pages were either less relevant or repetitive when used a different phrase from our list. We then sorted papers by their recentness and the reputation of their publication venues (e.g., recent papers in CHI, ASSETS came up first). Even though we enforced a restriction on the number of search results, we observed many duplicates. Next,

after removing duplicates, two authors manually reviewed each paper's title, abstract, and introduction in the sorted corpus. They additionally removed papers unrelated to remote assistance, sighted assistance, remote assistant systems, and navigation systems for people with VI. This process narrowed down the number of papers from 70 to 35.

The same two authors read those 35 papers thoroughly. Then, they further excluded 15 papers from the corpus because (i) those papers studied the usability of remote sighted assistance (RSA) system for image description, grocery shopping, and object detection; or (ii) those papers investigated the impact of RSA system in improving the quality of life of people with VI, discussed the privacy concern in RSA; or (iii) those papers focused on improving the interfaces or functionality of RSA system; or (iv) those papers focused on the feasibility of RSA system in navigation tasks (e.g., some papers used the word wayfinding or mobility) but did not reveal related challenges or requirements. Finally, the corpus included 20 papers for further analysis.

Three authors cross-checked the analysis to confirm that identified challenges are extracted from the literature and do not exceed the context. Table 1 summarizes these challenges. Below, we describe the relevant literature from where individual challenges are drawn.

3.1 Challenges in Localization and Orientation

One of the biggest challenges identified for the RSA agent is localizing the user and orienting the agent themselves. For this task, the agent mainly depend on the two sources of information - users' live video feed and GPS location. The agents put them together to localize the users on a digital map on their dashboard [16, 74]. However, the agents frequently got confused to perceive which direction the user is facing from the user's camera feed and GPS location [27, 43, 65]. The trained agents who participated in prior study [65] also reported losing users' current location is a hard problem. RSA agent's lack of environmental knowledge and their unfamiliarity of the area, scarce and limitation of the map, and inaccuracy of GPS found to be main causes for the location and orientation related challenges.

3.1.1 Unfamiliarity of environment. In previous study [73], the RSA agents expressed their frustrations with the users' expectation of the agents' quick start of assistance, which is usually not possible because most of places are new to the agents and thus they need some time to process the information to orient themselves. The fact that RSA agents' never being in the place physically but depending only on the limited map and the video feed is reported as a cause for the challenge in the following research work [43, 59, 65].

3.1.2 Scarcity and limitation of maps. Lee et al. [74] reported that RSA agents primarily use Google maps for outdoor spaces, and they perform Google search to find maps for indoor places. RSA agents who participated in Lee et al.'s study [74] reported that coarse or poor maps of malls or buildings limit their ability to assist the users. They also stated that many public places either have no maps or have maps with insufficient details, which forces them to rely on another sighted individual in close proximity of the user for assistance. Sometimes, agents must orient the users using their camera feeds only [43, 74], which makes the challenges worse. Navigating complex indoor layout is one of the well-established challenges in pedestrian navigation, as reported by many researchers [47, 65, 86, 94].

3.1.3 Inaccurate GPS. In addition to the insufficient map problem, inaccurate GPS was recognized as another major cause. Field trials of RSA system [16] revealed that the largest orientation and localization errors occurred in the vicinity of a tall university building where GPS was inaccurate. Researchers [43] indicated that GPS signal reception was degraded or even blocked around tall buildings. In terms of the last-few-meters navigation, they illustrated that GPS was not accurate enough to determine whether the user was walking on the pavement or the adjacent road in some situations. The well known last 10 meters and yard problem [101] in the blind navigation is also caused by the GPS inaccuracy.

3.2 Challenges in Obstacle and Surrounding Information Acquirement and Detection

The second notable challenges that the agents face occur in their obtaining the information of obstacles and surroundings. RSA agents need to detect obstacles vertically from ground level to head height, and horizontally along the body width [43, 62, 91]. They also need to provide information about dynamic obstacles (e.g., moving cars and pedestrians) and stationary ones (e.g., parked cars and tree branches) [16, 62]. However, agents found these tasks daunting due to the difficulties in estimating the distance and depth [65], reading signage and texts [59], and detecting/tracking moving objects [59, 62] from the users' camera feed. Number of research work also found that it is almost impossible for agents to project/estimate out-of-frame potential either moving or static obstacles [16, 27, 43, 62, 65, 91, 104]. Researchers [14] described that navigation environments in the real world are dynamic and ever-changing. Thus, it is easier for agents to detect obstacles and provide details when users are stationary or moving slowly [59]. Two main causes are linked with aforementioned problems: 1) limited field of view of the camera; 2) limitation of using video feed.

3.2.1 Narrow View of the Camera. Prior research found that the camera in use had a relatively limited viewing angle of around 50°, compare to the angle of human vision that is up to 180° [27]. Researchers [62, 104] mentioned that the camera should be located appropriately to maximize vision stability along the path. Limited field of camera view affects RSA negatively in their guiding performance [63, 65].

3.2.2 Limitation of using video feed. The quality of the video feed that matters to the RSA is the steadiness and clearness. The video stream easily affected by the motion of the camera (e.g., handheld mobile device or glasses) and becomes unstable. It is reported that agents are more likely to experience motion sickness when the users are not holding the camera (e.g., smartphone hanging around the user's neck) [59]. To mitigate the challenges of reading signage and texts in the user's camera feed, researchers [27] demonstrated the necessity of enhancing the quality of the video stream. Smooth frame rate and high resolution are essential when agents read signs, numbers, or names. The quality of the video stream can affect the performance of RSA in hazard recognition [41, 42], and thus it is considered as one of the main factors determining the safety of blind

Challenges in RSA

Problems	Causes	Needs for Design Space
G1. Orientation and Localization		
(1) Scarcity of indoor map [65, 74, 86, 94] (2) Unable to localize the user in the map in real time [16, 43, 74, 86] (3) Difficulty in orienting the user in his or her current surroundings [27, 43, 94] (4) Lack of landmarks or annotations on the map [14, 74] (5) Outdated landmarks on the map [14, 74] (6) Unable to change scale or resolution in indoor maps [74] (7) Last-few-meters navigation (e.g., Guiding the user to the final destination) [86, 101]	(1) RSA agent's lack of environmental knowledge (e.g., the agent is not familiar with the surroundings of the blind user) [43, 59, 65] (2) Scarcity of maps and Limitations of using maps (3) Inaccuracy of GPS	(1) Detailed map of indoor and outdoor (2) Improved interactivity of the map
G2. Obstacle and Surrounding Information Acquirement and Detection		
(1) Difficulty in reading signages and texts in the user's camera feed [59] (2) Difficulty in estimating the depth from the user's camera feed and in conveying distance information [65] (3) Difficulty in detecting and tracking moving objects (e.g., cars and pedestrians) [59, 62] (4) Unable to project or estimate out-of-frame objects, people, or obstacles from the user's camera feed [16, 27, 43, 62, 65, 91, 104] (5) Motion sickness due to unstable camera feed [59]	(1) Limitation of using video feed (e.g., unstable video stream) (2) Limited camera field of view; narrow view of camera	(1) Enhanced and improved camera field of view and video feed (2) Augmented visual information on video feed
G3. Delivering Information and Understanding User Specific Situation		
(1) Difficulty in proving various information (direction, obstacle, and surrounding) in timely manner [73, 74] (2) Adjusting the pace and level of detail in description provision through communication [59, 74] (3) Cognitive overload	(1) The need of delivery of large volume of information (2) Differences in preferences and various context/situation of users	(1) Information prioritization support (2) Cognitive load reducing interface and interaction design (3) Augmentation of information with different modality (4) Collaborative interaction between the agent and the users
G4. Network and External Issues		
(1) Losing connection and low quality of video feed [30, 43, 59, 62, 65, 74] (2) Poor quality of the video feed	(1) Weak signal and network in indoor and some places (2) Low ambient lighting	(1) Offline map (2) Camera feed enhancement

Table 1: A list of challenges in RSA service, presented in four groups (G1, G2, G3, and G4).

users [16]. This explains that the task of an intersection crossing was recognized as one of the most challenging situations for agents in a navigation aid. RSA agents find it very challenging because it is difficult to identify traffic flow through the narrow camera view, poor video quality, and the high speed of vehicles [27, 59, 65].

3.3 Challenges in Delivering Information and Interacting with Users

In addition to the challenges in the task of obtaining the necessary information, agents informed next set of challenges happened in delivering the obtained information and interacting with the users. In previous studies [73, 74], agents revealed the difficulties in providing various required information (e.g., direction, obstacle,

and surroundings) in timely manner and prioritizing them, which requires understanding and communicating with users that creates further challenges. The agents could also stress out if the users move faster than they could describe the environment [74]. These suggest that, in the navigation task, the need of delivery of large volume of information and simultaneously, the need of quick grasp of each user's different situation, need, and preference are main causes for the challenges. Prior research found that the RSA deal with this challenges through collaborative interaction/communication with the users [59, 74].

3.4 Network and External Issues

Early implementations of RSA services suffered from the network connection and limited cellular bandwidth [43]. Although cellular connection improved over the years, the problem remains for indoor navigation [59], which could lead to large delays or breakdowns of video transmissions [15, 62, 65]. Also an external factor such as low ambient light condition at night causes the poor quality of the video feed.

4 IDENTIFYING NAVIGATION CHALLENGES IN RSA: USER INTERVIEW STUDY

Next, we conducted a semi-structured interview study with 12 visually impaired RSA users to understand the navigational challenges from the perspective of RSA users' experience. In this section, we report the findings of the users' experienced challenges, their perceptions of RSA agents' challenges, and how the challenges on each side of RSA provider and users are related and affect the RSA navigation experience.

4.1 Participants: RSA Users with VI

We recruited a total of 12 participants: 8 through RSA service company, (Aira [8]) and 4 from our prior contacts. All participants were familiar with free (e.g., BeMyEyes [17]) and paid (e.g., Aira [8]) RSA services. Aira company advertised about our study to their customers with VI and those who were interested in our study directly contacted us for participation. Our interviewees (9 female, 3 male) have various levels of visual impairments, and their ages range from 19 years to 62 years old. On average, they used a paid RSA service for at least one year. They also use white canes. Their participation was voluntary, and no compensation was provided.

4.2 Procedure and Data Analysis

The interviews were semi-structured, performed remotely over phone calls, and lasted between 30 minutes to 60 minutes. The researcher took an open-ended approach and used the following questions as an anchor to probe the follow-up questions for a deeper understanding of the challenges and issues centered on the topic of navigation. The three categories of the questions included: (i) identification of common navigation scenarios and the reasons for the need of RSA service; (ii) challenges that RSA agents faced, as perceived by the RSA users (interviewees) and the strategies they used to help the agents; and (iii) scenarios that were challenging either for them or the agents or both. All interviews were audio-recorded with the consent of the interviewees and transcribed. With the transcribed data, we developed a deductive coding framework with core topics drawn from 3 areas of the interview questions, which we considered high-level categories (e.g., common navigation scenarios, challenges for RSA users and agents, strategies/help for the challenges). We used these categories to organize all the transcribed data. Afterward, 2 researchers independently performed an iterative inductive analysis [24] on the data and generated codes (e.g., navigating in a parking lot, finding a trash can), additional categories (e.g., indoor and outdoor scenarios), and themes (e.g., indoor and outdoor specific challenges, lack of maps, unfamiliarity of environment). All codes and themes were reviewed with the processes of merging and refining, and final themes were extracted.

4.3 Findings

From the interview study, we identified challenging indoor and outdoor navigational scenarios from the blind user's experience (Table 3). Further we saw that major problems recognized from the literature review (e.g., the limitations of maps, RSA's environmental knowledge, and the camera view and feed) reappear as the main causes for challenges of the blind users and found that how those problems affect the users' navigation experience, and how they perceive and help the problems on the users' end.

4.3.1 Common Navigation Scenarios. The most common types of navigation scenarios that our participants ask RSA agents for help are traveling and navigating unfamiliar indoor or outdoor places. Navigating unfamiliar areas where a blind user might utilize RSA service came up often in our study, which is consistent with literature [73, 74].

For outdoor navigation, common scenarios include checking and confirming the location after Uber or Lyft drop-offs; finding an entrance from a parking lot; taking a walk to a park, coffee shop, mailbox; navigating in a big college campus; and crossing street.

The common indoor places they called RSA agents for help were Airport and large buildings (e.g., malls, hotels, grocery stores, theaters). In an airport, they usually ask RSA agents to find a gate and baggage claim area. Inside large establishments or buildings, they ask for finding certain point-of-interest (e.g., shops, customer service desk); entrance and exit, stairs, escalator, and elevator; and objects, e.g., vending machine and trash can.

Our data suggest that blind users repeatedly use RSA services to navigate the same place if its layout is complex (e.g., airports); or their destination within the place is different (e.g., different stores in a shopping mall); or the place is crowded and busy (e.g., restaurants).

4.3.2 Challenging Outdoor Navigation Experiences. It was a recurrent theme that if the agents experience challenges, the users also experience challenges. The interviewees were mostly content with their outdoor navigation experience with the agent, compared to that of indoor navigation, even though they realized that some scenarios were challenging to agents. Examples of such scenarios include crossing intersections, and finding certain places and locations (e.g., building entrances, restrooms) in open outdoor spaces (e.g., parking lots, campus).

Christi commented about the challenge in parking lots: *"Parking lots aren't fun, even with an Aira agent unless there is a walking path"*. Denise also shared her experience of taking longer than usual time to find a public restroom in a big open bus stop. She stated that

Pseudo name	Gender	Age	Age of Onset	Condition of Vision Impairment	Occupation
Calvin	M	32	4yrs	Legally blind, 20/700, shaky eyes	Student (Computer science major) /consulting for family and friends, beta testing app
Grace	F	19	At birth	Total blindness, a little bit of light perception	Student (special education/ vision major - dual major)
Karen	F	23	12yrs	Usable vision in right eye color, shape	Student (liberal studies major)
Larry	M	62	39yrs	Total blindness	Peer support specialist library social worker (scholarship plan)
Susan	F	27	At birth	Glaucoma; light perception on both eyes	Student/technology instructor
Sally	F	25	At birth	Total blindness	Student (family studies)
Justin	M	45	6yrs	Total blindness	Software tester
Christi	F	44	At birth	Total blindness, no light perception	Social worker
Denise	F	55	At birth	Total blindness	Counselor
Hanna	F	35	At birth	Total blindness	Student/technical advisor
Kelly	F	32	At birth	Premature, congenital	Student (political science major)
Rachel	F	22	At birth	Total blindness	Student (Special education)

Table 2: Participants' demographics in our study with RSA users.

frequent incidences of getting to locked doors were very unpleasant experiences: *"Probably the most challenging is if I'm outdoors at a location and I'm trying to find the door to go into, some doors are locked"*. Larry shared his experience of having incorrect guidance from the agent that led him to the wrong place. He attributed it to the mismatch of the map and the real place.

> *"This kind of frustrated me...for some reason, it wasn't matching up with what they[agent] were seeing on their map. And I guess it wasn't matching up with what they were visually seeing. The agent took me on a whole different path." - Larry*

4.3.3 Challenging Indoor Navigation Experiences. All our interviewees commonly mentioned that indoor navigation was more challenging for them, as well as for the agents. Sally clearly stated that: *"again, it was that indoor navigation that gave them [Aira agents]...an issue"*. Interviewees' indoor experiences with RSA indicate that it usually takes much longer for the agent to locate and find in indoor spaces.

Christi shared her challenging experience of spending about 20 min with a RSA agent only to find a store ("Bath and Body Works") from another store in a big mall. She recalled they both got lost and disoriented in "JC Penny":

> *"She [Aira agent] could not get us navigated out of there at all, like through the store to get to the other part of the mall." - Christi*

Another interviewee, Rachel, recounted the longest and the most challenging time she had with an agent when trying to find the luggage section in a department store, Kohls.

> *"Eventually [I] locate it but I know I was walking around in the store [for] a long time ... a lot of back and forth across the store." - Rachel*

Finding the entrance (or exit) of a building, navigating to a pick-up point from the interior of a building for meeting the ride-sharing driver are examples of other challenging experiences that our interviewees shared with the RSA agents. Denise explained what happened in a big grocery store:

> *"Kroger has different places that you can go in or out, and the Uber or Lyft, whatever ride I was taking, let me out at one entrance, but then they came to pick me up at a different entrance. ... They did not know which entrance I came in and which one I was going to go out of, so that was a challenge." - Denise*

She added frequent incidences of getting to locked doors was very unpleasant experience: *"Probably the most challenging is if I'm outdoors at a location and I'm trying to find the door to go into, some doors are locked"*. Also, finding a seat in a restaurant or a theatre looked challenging tasks for RSA to RSA user interviewees.

4.3.4 Users' Understanding of Problems: Insufficient Maps, RSA's Unfamiliarity of Area, and Limited Camera View. All interviewees mentioned that the absence of maps or floorplan, as well as the inaccuracy and scarcity in the map, are the primary reasons why RSA agents struggled to assist them in both indoor and outdoor places. Justin pointed out this lack of map issue:

> *"... The second biggest problem is not being able to find good maps... Like if I'm in an airport, they can't always find a good map of the airport." - Justin*

Most interviewees additionally mentioned that the RSA agent's unfamiliarity with a place and location is an obvious challenger. Susan and Larry described this challenge as follows:

> *" when you're walking to a new place, the agent doesn't know the campus. ..they have the map and they have their visual cues from what they're looking at from your*

> *end, from your glasses [Aira glasses] or your phone. But, you know, it's not like they know the campus. It's obviously,..it's a matter of getting those routes down and that takes time, because you had to walk in one direction, walk in another, just to kind of scope out the route, the one throughout is scoped out." - Susan*

> *"...I am in a store or in a hospital, you know, they don't know, where they're,...they have Google Maps, you know, that's it. It is a lot of back and forth. But I find that...that be the case for every, essentially, every new place." - Larry*

Several interviewees' accounts also suggested that poor and limited visibility caused by the narrow camera view creates challenges for the agents. Karen talked about using stairs in her school and showed her understanding of the agent's visual challenge caused by the limited distance that camera feed can show.

> *"Going down flights of stairs can be a challenge because the agent can't see that far down, but that's when I totally understand why they say to have our cane at all times." - Karen*

She introduced another visual challenge that the current camera's capability can't solve and Calvin's comment implies the same issue.

> *"my college campus we have a lot of stairs that are very narrow, they're very like close together and so it's hard to tell that they're actually a staircase so sometimes the agents don't see that there's actually a set of stairs there because the stairs are so close together." - Karen*

> *"... not sure if [the] sign was blocked or hidden..." - Calvin*

4.3.5 Users' Helping and Collaborating with RSA. The participants seem to understand the difficulties and situations that the agents face on their ends and they want to assist the agents and willing to work with them to mitigate the challenges and complete the navigation task together.

The story of how Susan collaboratively worked with the agent and created a map that accommodates her specific needs and how it helps both her and the agent every time she is connected with the agent was impressive and intriguing.

> *"I recreated the map. the map was on Google, and then instructions for the map. So in the Aira folder, I have a map of my college. So I have routes in there that Aira and I have done, Aira agents and I have done together. There have been instructions for each route, so they've marked off the map, and they've also written instructions on with the route... they can see exactly where I am...they quite easily get me on the path where I needed to go. So that's been really helpful on getting me around the campus." - Susan*

Karen and Grace shared what she usually do to help the agent get a better view of the video feed for the direction and distance on their end. Karen said She pays attention to positioning the phone in the right way so the agent can see where they need to see.

Grace tried to enhance the view by adjusting the distance.

> *"When I'm calling an agent, I would hold the phone so that my finger isn't blocking the camera and try to keep it kind of farther away from the object so they can get a better view of it." - Grace*

The participants also mentioned that a common workaround that RSA agents use in challenging scenarios is to find a dependable sighted person, such as an employee with a uniform or someone in the helpdesk, who could help them quickly. A similar finding was also reported by Lee et al. [74].

5 DISCUSSION: ADDRESSING RSA CHALLENGES

The findings showed that the interviews with the blind RSA users echoed the challenges and difficulties identified from the literature review. In this section, we discuss the potential of the existing CV technology for addressing those challenges, specifically with the 3D map created through a framework of human-AI collaboration and how much of the problems could be mitigated with those potential technological solutions. In the following sections, we present how existing CV-based techniques can be used to address a subset of challenges listed in Table 1.

A fundamental challenge for RSA is the lack of situation awareness for the user's surroundings. Traditionally, this is mitigated by providing the sighted agents with a live camera feed and a map (e.g., *Google Maps*). However, these traditional means have shortcomings. Further, navigation is a teamwork-based task that demands close collaboration between users and RSA agents as found in the prior work [74] and also in our user interview study presented above. In theory, representing the real world in 3D digital maps can address challenges like lack of indoor maps and agent's lack of environmental knowledge. If such maps are interactive, allow collaborative annotation, and support map-based localization and path planning, these can address the challenges caused by the limited view of camera [120]. Other developments, such as augmenting video streams with texts, graphics, and detected objects, can address challenges like difficulty in conveying distance information, reading signage and text in a video feed, highlighting landmark, and reducing agents' context switches.

5.1 3D Map Construction for the RSA Use

3D mapping is the profiling of real-world objects in 3D space. Compared to conventional 2D maps, 3D maps provide a more realistic and intuitive view of the environment. A 3D map is usually represented by point cloud data, which comprise a dense set of vertices in 3D space. Point clouds can be captured by various sensors, including Light Detection and Ranging (LiDAR), Time-of-Flight (ToF) cameras, and RGB cameras with photogrammetry techniques. We can use smartphones' built-in augmented reality (AR) frameworks, such as ARKit [3], ARCore [1], to generate point clouds. Besides, the new iPhone 12 Pro and iPad Pro also integrate a LiDAR scanner, making it possible to create high-quality 3D maps directly from hand-held devices. The point clouds generated by AR frameworks are much denser than those created from RGB videos with structure from motion (SfM) pipelines [45]. Depending on the application scenarios, mapping the locations of interest can be implemented

Challenging Scenarios for Users

Outdoor scenarios	Indoor scenarios
1. Going to mailbox	1. Finding trash cans or vending machines
2. Taking a walk around a familiar area (e.g., park, campus)	2. Finding architectural features (e.g., stairs, elevators, doors, exits, or washrooms)
3. Walking to the closest coffee shop	3. Finding a point-of-interest in indoor navigation (e.g., a room number, an office)
4. Finding the bus stop	4*. Navigating malls, hotels, conference venues, or similarly large establishments
5*. Crossing noisy intersections without veering	5. Finding the correct train platform
6. Calling a ride-share and going to the pick-up location	6. Navigating airport (e.g., security to gate, gate to gate, or gate to baggage claim)
7*. Navigating from a parking lot or drop-off point to the interior of a business	7. Finding an empty seat in theaters or an empty table in restaurants
8*. Navigating through parking lots or construction sites	

Table 3: All 15 scenarios were reported by all participants. Scenarios with $*$ occurred more frequently than others. Also, participants perceived these as more challenging than others.

in either an offline, crowd-sourcing manner or an on-the-fly manner. With 3D indoor maps, the problem G1.(1) in Table 1 can be addressed.

The possibility of creating 3D map from an ARKit built-in handheld device using a smartphone application shows a great potential not only for addressing major problems in the Challenges in RSA category, *Orientation and Localization* but also for providing a platform that can enable the multiple levels of human-AI collaboration (RSA agent-AI-RSA user). As presented in section 2, the feasibility and effectiveness of the technology itself has been evaluated and demonstrated by other researchers' investigations [114, 123] with the interest of directly helping people with VI, not the RSA agents. However, considering its newness and known inaccuracy issues [101] that could create undesirable experiences for users with visual impairment as well as the potential benefits for the RSA's notable problems, it is opportune to investigate the use of technology in the RSA system.

The user App can be equipped with ARKit or ARCore framework. Using this app, sighted volunteers can iteratively construct offline maps by carefully scanning interesting areas. During scanning, the generated point clouds are uploaded to a computer server in real-time. Later, when another sighted volunteer scans the same area or building, the server automatically detects the correspondence between the current map and previously stored ones. If adequate overlapping areas are recognized, the current map will be merged to the matched map.

If offline 3D maps are not available, especially when an RSA user navigates a new environment, the App still can generate a new 3D map in real time for the areas recorded by the user's smartphone camera. Note that it stores 3D visual information of the whole area that the user's camera has scanned so far and continues to expand as the user moves, whereas the video feed only provides the current view in front of the camera. Compared to the video feed alone, the on-the-fly 3D maps provide a more holistic view of the scene and the user's movement. In a sense, the 3D map can serve as the "visual memory" for the agent.

5.2 Collaborative Annotation for Limitation of Map

The construction and use of the 3D map in RSA system can also allow sighted volunteers to interact with the system and annotate the 3D map in a collaborative way. This would enrich the map with the information needed for RSA agents, and thus address the identified problem G1.(4) and G1.(5) in Table 1. The generated 3D maps are point clouds without semantic information. To facilitate navigation task, sighted volunteers can manually annotate landmarks or objects on the 3D maps. Similar to map construction, this can be done in either an offline or online manner.

For offline annotation, sighted volunteers can create, edit, or delete landmarks (*e.g.*, meeting rooms, offices) on the point cloud data based on their knowledge of the buildings. Providing an additional tool and feature (e.g., search bar) along with the annotation functionality would make the edit process more efficient. The annotations are thus flexible for the dynamic environmental changes. One advantage of offline annotation is that the map construction and annotation steps are separated so that the respective workflows do not interrupt each other. During online annotation, sighted volunteers create, edit, or delete labels directly on the video feed. For example, while scanning the hallway, the sighted volunteer adds an annotation for the office number. This annotation is automatically transferred to the 3D map to minimize the interruption of mapping workflow.

5.3 Map-based Localization for Real-Time Locating and Tracking the User

Once the 3D map of a certain building has been constructed and annotated, the first step of using such a map would involve localizing the user on the map. This localization capability would compensate for the RSA agents' lack of environmental knowledge and unfamiliarity and the inaccuracy of GPS that create the problem of difficulty in orienting the users in their current surroundings. We imagine the dashboard that RSA agents look at would have divided windows: one side shows the indoor 3D map of the building in which the user

is currently located; the other side shows the live video feed from the user's smartphone.

When the user enters the building, the offline-built 3D map will be automatically loaded to the center window in the dashboard. Feature points are detected from the frames of the live video feed and matched with the 3D point cloud. After finding the initial 2D-3D correspondences between the images and the 3D map, feature points are continuously tracked in real-time. If the camera is calibrated, its real-time location and orientation can be computed from the 2D-3D correspondences by solving the Perspective-n-Point problem [57]. Because the user holds the smart device in a front-facing manner, the location and orientation of the user are the same as that of the camera. Thus, the system can show the user's location and orientation on the 3D map, addressing problem G1.(2), G1.(3), and G1.(7) in Table 1.

5.4 Interacting with 3D Maps for Efficient Information Retrieval and Delivery

Once the detailed 3D map with the annotated information is ready for the RSA agent, it would be greatly helpful if the agent could retrieve and view the only information she or he wants and needs on the map and customize the way the map is shown. This flexibility and interactivity with the map and accompanying information can address the identified problem G1.(6) and group G3 in Table 1.

With the user localized on the map, RSA agents interact with the map the same way they interact with 2D maps. For example, the agent can browse or search the landmark, and change the scale and viewpoint via zoom in/out, translation, and rotation. Sometimes, switching between views or returning to the top global view requires multiple steps and tweaks (zoom in/out, translation, and rotation). To make the interaction more efficient, the dashboard can contain several shortcut buttons to reset view, switch between the global map, current location first-person view, and destination.

Relying on either the annotated landmarks or online map exploration through an interactive interface, an RSA agent can find and mark the destination on the 3D map and perform a path planning. This additional interactive feature that helps determining the path in the beginning of interacting with the user can lessen the problem of the cognitive overload for the RSA agent (*i.e.*, problem group G3 in Table 1.) When the user's location and the destination are known, the RSA agent can further decide and draw a walkable path on the 3D map. Alternatively, the path can be automatically planned by an A^* path search algorithm [56] (similar to routes in Google Maps).

5.5 Augmenting Video Stream

If the planned path drawn on the 3D map also can be projected in the live video feed, it would make the RSA agent's user guiding task much easier. The current Augmented Reality (AR) technology can make this possible. Any real-time key information such as turn-by-turn directions, distances, and landmarks can be augmented in the live video stream. This additional AR powered presentation of the information would provide the RSA agent with a next level of information acquirement and process, and thus will help with the challenges in delivering information to the users.

It also can show distance bands of the user to points in the scene (e.g., 5 ft and 10 ft) based on the real-time localization on the 3D map. In addition, the annotated landmarks and objects in the 3D map can also be projected to the video feed with the appropriate distance information, e.g., *Office 101* and *Display Cabinet*. To facilitate measuring distances in a broader range, the global top-view of the map can be divided into grids, commonly used in cartography. With distance bands and grids, the problem G2.(2) in Table 1 can be addressed.

In the live video feed, the RSA agent may miss important landmarks or signs. The recent capabilities of computer vision techniques to detect objects and read scene texts can be leveraged to create additional AR labels on the live video [77, 126]. The text recognition function, as demonstrated by SeeingAI [82] app, can be particularly useful in challenging situations such as blurry or partially-occluded signs. At any time, RSA agents can enable or disable an augmented element.

6 DISCUSSION: EMERGING PROBLEMS IN HUMAN-AI COLLABORATION

In the following sections, we present a subset of challenges listed in Table 1 that cannot be addressed by existing CV-based techniques, for which we formulate five emerging human-AI collaborative approaches.

6.1 Emerging Problem 1: Making Object Detection and Obstacle Avoidance Algorithms Blind-aware

Obstacle avoidance is a main task in people with VI's navigation due to safety concerns. As discussed in literature review, detecting obstacles is a notable challenge through the narrow camera view, because the obstacle could appear vertically from ground level to head height, and horizontally along the body width [43, 62, 91], as listed in problem G3.(1) in Table 1. This requires the agents to observe the obstacles at a distance from the camera feed. But it is still extremely difficult for the agents because the obstacles afar would be too small to recognize in the camera feed. The challenge motivates us to resort to AI-based object detection algorithms [124], which are able to detect small objects. However, it is problematic to directly apply existing object detection algorithms [96, 97] to the RSA services. For example, a wall boarding a sidewalk is considered as obstacles in common recognition models but can be regarded as Orientation & Mobility (O&M) affordances for people with VI who use a cane and employ the wall as a physical reference. We term the ability of recognizing affordances that are important for people with VI as *blind-aware*, a common philosophy in end-user development [36]. Due to the importance of detecting obstacle in a blind-aware manner, we consider it as an emerging research problem that can be addressed by human-AI collaboration.

In the context of navigation, researches have adopted machine learning algorithms to automatically detect and assess pedestrian infrastructure using online map imagery (e.g., satellite photos [6, 7], streetscape panoramas [54, 55, 108]). Recent work [118] applied ResNet [58] to detect the accessibility features (e.g., missing curb ramps, surface problems, sidewalk obstructions) by annotating a dataset of 58,034 images from Google Street View (GSV) panoramas.

We can extend these lines of work to a broader research problem of detecting objects including accessibility cues in navigation. First,

we need volunteers to collected relevant data from satellite photos (e.g., Google Street, Open-street maps), panoramic streetscape imagery, 3D point clouds, and camera feeds of the users. Following [118], data-driven deep learning models are trained with human annotated data. It is worth noting that the data are not limited to images but also 3D mesh or point clouds, especially considering iPhone 12/13 Pro has equipped with LiDAR scanner. To train blind-aware models for object detection, we also need to manually define whether an object is blind-aware with the help of people with VI. Specifically, blind users can provide feedback on the quality of a physical cue [119]. Besides, another human-AI collaboration direction is to online update the computer vision (e.g., obstacle detection) models with new navigation data marked by the agents. Solving this problem could make blind navigation more customized to how the blind user navigates through space and expedite the development of the automated navigation guidance system for blind users.

6.2 Emerging Problem 2: Localizing Users Under Poor Networks

Although cellular bandwidth has been increased over the years, the bad cellular connection is still a major problem in RSA services, especially in indoor navigation [59], as listed in problem G4.(1) in Table 1. The common consequences include large delays or breakdowns of video transmissions [15, 62, 65]. Suppose the poor network only allows transmitting limited amount of data and cannot support live camera feed, it is almost impossible for the agents to localize the user and give correct navigational instructions. Based on this observation, we identify an emerging research problem of localizing users under poor networks that can be addressed by human-AI collaboration.

With regard to AI-based methods, one possible solution is to use interactive 3D maps, constructed with ARKit [3] using an iPad with a LiDAR scanner. During an RSA session under a poor network, the user's camera can relocalize them in the 3D maps. If their location and camera pose is transmitted to the agents, agents can simulate their surroundings on the preloaded offline 3D maps. Considering the camera pose can be represented by a 4×4 homogeneous matrix, the transmitted data size is negligible. With voice chat and the camera pose displayed on the 3D maps, the agent can learn enough information about the user's surroundings and localize the user under a poor network momentarily.

In terms of human-AI collaboration, to the best of our knowledge, there is no work for RSA on localizing users under poor networks. Without live camera feed, it would be a more interesting human-AI collaboration problem. To localize the user in such situation, the communication between the agent and the user would be greatly different. We can imagine some basic communication patterns. First, the agent can ask the user to make certain motions (e.g., turn right, go forward) to verify the correctness of the camera pose display. In turn, the user can actively ask the agent to confirm the existence of an O&M cue (e.g., a wall) from the 3D maps. It is worth noting that the offline 3D map could be different from the user's current surroundings. When exploring the map, they also need to work together to eliminate the distraction of dynamic objects (e.g., moving obstacles) which do not exist on the 3D map. The detailed problems have never been studied. For example, how to detect the localization errors and maintain the effective RSA services in low data transmission rate.

6.3 Emerging Problem 3: Recognizing Digital Content on Digital Displays

Digital displays, such as LCD screens and signages, are widely used in everyday life to present important information, e.g., flight information display board at the airport, digital signage at theaters, and temperature control panel in the hotel. RSA agents reported difficulty in reading texts on these screens when streamed through the users' camera feed, as listed in problem G2.(1) in Table 1. This difficulty can be caused by several technical factors, including varying brightness of a screen, i.e., the display of a screen is a mixture of several light sources, e.g., LCD backlight, sunlight, lamplight [89]; a mismatch in the camera's frame rate and the screen's refresh rate; and a mismatch in the dimension of pixel grids of the camera and the screen, resulting in moiré patterns, i.e., showing strobe or striping optical effects [89]. Based on the significance and challenges of recognizing content on digital displays through camera feeds, we consider it as an emerging research problem that can be addressed by human-AI collaboration.

From the perspective of AI solutions, there exist a few computer vision systems that assist blind users to read the LCD panels on appliances [38, 49, 85, 111]. However, these systems are heuristic-driven, fairly brittle, and only work in limited circumstances. To the best of our knowledge, there is no text recognition method specifically designed to recognize digital texts on LCD screens or signages in the wild.

In this regard, we consider scene text detection and recognition [77] as the closest computer vision method aiming to read texts in the wild. However, these methods are far more difficult than the traditional optical character recognition (OCR) of texts from documents. For example, the state-of-the-art deep learning methods [18, 116, 122] only achieve $< 85\%$ recognition accuracy on the benchmark dataset ICDAR 2015 (IC15) [66]. Furthermore, existing methods for scene text recognition are likely to suffer from the domain shift problem due to the distinct lighting condition [115], resulting in even worse recognition performance in reading digital content on LCD screens.

To formulate human-AI collaboration, we consider scene text recognition methods [77] as the basis for AI models. Next, we consider three aspects of human-AI collaboration. *First*, computer vision techniques can be used to enhance the camera feed display [12], while the agents are responsible for the content recognition. In this way, the content in the live camera feed will be transferred to have better lighting and contrast, making them more suitable for the agents to perceive and recognize.

Second, scene text recognition methods [77] can be used to read the digital content for the agents and provide the recognition confidence. It may be especially useful in recognizing small-scale text which is too small in the camera display for the agents to read but with enough pixels for the AI models to process. The agent can ask the user to change the camera angle to get a better view to achieve better recognition results.

Third, the agents are usually interested in recognizing certain texts on the screen, thus can mark the region of interest for AI

to process. In this manner, the agents can improve the processing speed of AI models, as well as reduce the models' unwanted, distracting outputs.

Note that the above three aspects of human-AI collaboration overlap, e.g., the enhanced camera feed can be used for both humans and AI to recognize better. Since it is still an open problem, there may be other aspects of human-AI collaboration to explore in the future. For example, to train AI models specifically for digital text on LCD screens, we need volunteers to collect pictures of digital content on LCD screens or signages from different sources (e.g., Internet, self-taken) with various conditions (e.g., image resolution, character size, brightness, blurriness) and annotate the location and content of the text on the pictures. VizWiz [51] dataset has set one such precedent. This dataset contains over 31,000 visual questions originating from blind users who captured a picture using their smartphone and recorded a spoken question about it, together with 10 crowdsourced answers per visual question.

6.4 Emerging Problem 4: Recognizing Texts on Irregular Surfaces

Reading important information on irregular surfaces (e.g., curved surface, non-orthogonal orientation) is common in people with VI's lives, e.g., reading the instructions on medical bottles, checking the ingredients on packaged snacks or drink bottles. However, it is extremely challenging for the agents to recognize text on irregular surfaces through the camera feed [59] due to the distorted text and unwanted light reflection, as listed in problem G2.(1) in Table 1. Therefore, we identify an emerging research problem of reading text on irregular surfaces that can be addressed by human-AI collaboration.

As far as only AI techniques are considered, scene text detection and recognition methods [77] could offer possible solutions to this problem based on the discussions in Problem 3. But the weaknesses of the pure AI solutions are similar to that in Problem 3. First, the state-of-the-art scene text recognition methods [18, 116, 122] still cannot perform satisfactorily on benchmark datasets. Second, existing text recognition methods [77] mostly read text on flat surfaces, and there are no methods specifically designed for recognizing text on irregular surfaces. When directly applying existing methods to reading text on irregular surfaces, the recognition accuracy would degrade further owing to the text distortion and light reflection.

Without regard to human-AI collaboration, scene text recognition methods [77] read text only relying on the trained AI models but not considering human inputs, while existing RSA services take no account of the potential applications of AI-based methods. Similar to Problem 3, we consider three main aspects of human-AI collaboration in recognizing text on irregular surfaces. *First*, the computer vision techniques can rectify the irregular content [105] and augment the video (e.g., with super-resolution [117]), and the agents recognize the text from the augmented video. *Second*, the agents can ask the user to move/rotate the object (e.g., medicine bottle) or change the camera angle to have a better view, and the AI models [77] can help recognizing the text, especially the small characters. *Third*, the agents select the region of interest on the irregular surfaces in the video for AI to process by either augmenting display or recognizing text. In addition, volunteers may be needed to collect images of text on different irregular surfaces (e.g., round bottles, packaged snacks) with various conditions (e.g., image resolution, character size, viewing angle) and annotate them for training customized AI models.

Despite similarities, there are three main differences between Problem 3 and Problem 4: (i) Problem 3 addresses the text recognition problem for luminous digital screen, but Problem 4 focuses on the text on non-luminous physical objects; (ii), the text in Problem 3 is on planar screens, but Problems 4 address the recognition on irregular (e.g., curved) surfaces. Thus, they require different customized AI models; and (iii) the screens in Problem 3 are usually fixed, and the user can move the camera to get a better view angle. In contrast, the objects with text in Problem 4 are movable. For example, the user can rotate the medicine bottle as well as changing the camera angle. That is, Problem 4 supports more interaction patterns than Problem 3.

6.5 Emerging Problem 5: Predicting the Trajectories of Out-of-Frame Pedestrians or Objects

In RSA services, the agents need to provide the environmental information in the user's surrounding (e.g., obstacles and pedestrian dynamics) for safety when the user is in a crowded scene. The trajectory prediction of pedestrians or moving objects could assist the agent to provide timely instructions to avoid collision. According to the literature review, it is extremely difficult for the RSA agents to track other pedestrians/objects [59, 62] from the users' camera feed, and almost impossible to predict the trajectories of out-of-frame pedestrians or objects [16, 27, 43, 62, 65, 91, 104], as listed in problem G2.(3) and G2.(4) in Table 1. The main reasons are the narrow view of the camera and the difficulty of estimating the distance. Based on this observation, we pose an emerging research problem of predicting the trajectories of out-of-frame pedestrians or objects that can be addressed by human-AI collaboration.

If only considering AI solutions, we can adopt human trajectory prediction technology [100] which has be studied as a computer vision and robotics problem. Specifically, the motion pattern of pedestrians/objects can be learned by a data-driven behavior model (e.g., deep neural networks). Then, based on the observation from the past trajectories, the behavior model can predict the future trajectories of the observed pedestrians/objects. There are two types of problem settings, i.e., observed from either static surveillance cameras [9, 50] or moving (hand-held or vehicle-mounted) cameras [79, 121]. For RSA application, we focus on predicting from hand-held cameras. Existing trajectory prediction methods forecast the future pixel-wise locations of the pedestrians on the camera feed without considering the out-of-frame cases. The pixel-level prediction is also not useful for the agents to estimate the distance to avoid collision. Moreover, existing models are learned from the scene without people with VI, but the motion patterns of pedestrians around people with VI could be rather different.

In terms of human-AI collaboration, to the best of our knowledge, there is no work exploring the problem of pedestrian tracking and trajectory prediction under active camera controls. We consider three aspects of human-AI collaboration in predicting the trajectories of out-of-frame pedestrians. *First*, we need to develop

user-centered trajectory prediction technologies. On one hand, the behavior models need to be trained from a people with VI-centered scene. On the other hand, the predicted trajectories should be projected to the real world where even the pedestrians cannot be observed from the camera feed. Based on such trajectory predictions, the agents can quickly plan the path and provide instructions to the user. *Second*, the agents may be only interested in the pedestrian dynamics towards the user's destination. In this case, the agents can mark the region of interest for AI models to conduct prediction. Then, AI models will save some computational resources and also understand the interest of the agents. *Third*, in turn, AI models could suggest moving the camera towards a certain direction (e.g., left) to get more observations for better predictions. In this way, AI models can better reconstruct the scene for the agents to make navigational decisions for the user. This problem can be further extended in the human-AI collaboration setting. For example, AI could offer suggestions on the user's walking directions with motion planning algorithms [84] based on the prediction results.

7 CONCLUSION

We first synthesize an exhaustive list of navigational challenges in agent-user interaction in RSA services through a literature review and a study with 12 visually impaired RSA users. Next, drawing on the prior work on computer vision-mediated RSA service, our analysis shows that some identified challenges cannot be addressed by off-the-shelf computer vision techniques because of the complexity of the underlying problems. Finally, we envision that these challenges can be addressed by the collaboration between RSA agents and computer vision systems. Therefore, we formulate five such emerging human-AI collaboration problems in the context of computer vision-mediated remote-sighted assistance. We hope our problem formulation will inspire researchers working in this area to take on these problems and open up new opportunities to enhance the RSA assistive experience.

ACKNOWLEDGMENTS

This research was supported by the US National Institutes of Heath, National Library of Medicine (R01 LM013330).

REFERENCES

[1] 2021. ARCore. Retrieved June 27, 2021 from https://developers.google.com/ar
[2] 2021. Autour. http://autour.mcgill.ca/en/.
[3] 2021. More to Explore with ARKit 5. Retrieved June 27, 2021 from https://developer.apple.com/augmented-reality/arkit/
[4] 2021. OpenStreetMap. https://www.openstreetmap.org/.
[5] 2021. The Seeing Eye GPS™ App in the iTunes Apple Store! http://www.senderogroup.com/products/shopseeingeyegps.html.
[6] Dragan Ahmetovic, Roberto Manduchi, James M Coughlan, and Sergio Mascetti. 2015. Zebra crossing spotter: Automatic population of spatial databases for increased safety of blind travelers. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. 251–258.
[7] Dragan Ahmetovic, Roberto Manduchi, James M Coughlan, and Sergio Mascetti. 2017. Mind your crossings: Mining GIS imagery for crosswalk localization. *ACM Transactions on Accessible Computing (TACCESS)* 9, 4 (2017), 1–25.
[8] Aira. 2021. Aira. https://aira.io/.
[9] 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.
[10] 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.
[11] 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.
[12] Mohamad Nurfakhrian Aziz, Tito Waluyo Purboyo, and Anggunmeka Luhur Prasasti. 2017. A survey on the implementation of image enhancement. *Int. J. Appl. Eng. Res* 12, 21 (2017), 11451–11459.
[13] 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.
[14] Nikola Banovic, Rachel L Franz, Khai N Truong, Jennifer Mankoff, and Anind K Dey. 2013. Uncovering information needs for independent spatial learning for users who are visually impaired. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.
[15] Przemyslaw Baranski, Maciej Polanczyk, and Pawel Strumillo. 2010. A remote guidance system for the blind. In *The 12th IEEE International Conference on e-Health Networking, Applications and Services*. IEEE, 386–390.
[16] 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.
[17] BeMyEyes. 2021. Be My Eyes. https://www.bemyeyes.com/.
[18] Ayan Kumar Bhunia, Aneeshan Sain, Amandeep Kumar, Shuvozit Ghose, Pinaki Nath Chowdhury, and Yi-Zhe Song. 2021. Joint Visual Semantic Reasoning: Multi-Stage Decoder for Text Recognition. *arXiv preprint arXiv:2107.12090* (2021).
[19] 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.
[20] BlindSquare. 2020. BlindSquare iOS Application. https://www.blindsquare.com/.
[21] 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.
[22] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. 2013. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2117–2126.
[23] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual recognition with humans in the loop. In *European Conference on Computer Vision*. Springer, 438–451.
[24] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] 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
[30] 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.
[31] 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.
[32] 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
[33] Mostafa Elgendy, Miklós Herperger, Tibor Guzsvinecz, and Cecilia Sik Lanyi. 2019. Indoor Navigation for People with Visual Impairment using Augmented Reality Markers. In *The 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 425–430.
[34] 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).
[35] 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.

[36] G. Fischer, E. Giaccardi, Y. Ye, A. G. Sutcliffe, and N. Mehandjiev. 2004. Meta-Design: A Manifesto for End-User Development. *Commun. ACM* 47, 9 (Sept. 2004), 33–37. https://doi.org/10.1145/1015864.1015884

[37] 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.

[38] Giovanni Fusco, Ender Tekin, Richard E Ladner, and James M Coughlan. 2014. Using computer vision to access appliance displays. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 281–282.

[39] 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.

[40] 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.

[41] Vanja Garaj, Ziad Hunaiti, and Wamadeva Balachandran. 2007. The effects of video image frame rate on the environmental hazards recognition performance in using remote vision to navigate visually impaired pedestrians. In *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology*. 207–213.

[42] Vanja Garaj, Ziad Hunaiti, and Wamadeva Balachandran. 2010. Using remote vision: the effects of video image frame rate on visual object recognition performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40, 4 (2010), 698–707.

[43] 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.

[44] Cole Gleason, Dragan Ahmetovic, Saiph Savage, Carlos Toxtli, Carl Posthuma, Chieko Asakawa, Kris M Kitani, and Jeffrey P Bigham. 2018. Crowdsourcing the installation and maintenance of indoor localization infrastructure to support blind navigation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–25.

[45] Cole Gleason, Anhong Guo, Gierad Laput, Kris Makoto Kitani, and Jeffrey P. Bigham. 2016. VizMap: Accessible Visual Information Through Crowdsourced Map Reconstruction. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*. ACM, 273–274.

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

[47] João Guerreiro, Dragan Ahmetovic, Daisuke Sato, Kris Kitani, and Chieko Asakawa. 2019. Airport accessibility and navigation assistance for people with visual impairments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[48] João Guerreiro, Daisuke Sato, Saki Asakawa, Huixu Dong, Kris M Kitani, and Chieko Asakawa. 2019. CaBot: Designing and evaluating an autonomous navigation robot for blind people. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 68–82.

[49] Anhong Guo, Junhan Kong, Michael Rivera, Frank F Xu, and Jeffrey P Bigham. 2019. Statelens: A reverse engineering solution for making existing dynamic touchscreens accessible. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 371–385.

[50] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2255–2264.

[51] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.

[52] 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.

[53] 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.

[54] Kotaro Hara, Jin Sun, Jonah Chazan, David Jacobs, and Jon E Froehlich. 2013. An initial study of automatic curb ramp detection with crowdsourced verification using google street view images. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[55] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. 2014. Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 189–204.

[56] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern.* 4, 2 (1968), 100–107.

[57] R. Hartley and A. Zisserman. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[59] 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.

[60] Bill Holton. 2015. Crowdviz: Remote video assistance on your iphone. *AFB AccessWorld Magazine* (2015).

[61] Bill Holton. 2016. BeSpecular: A new remote assistant service. *Access World Magazine* 17, 7 (2016).

[62] 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.

[63] Ziad Hunaiti, Vanja Garaj, Wamadeva Balachandran, and Franjo Cecelja. 2005. Use of remote vision in navigation of visually impaired pedestrians. In *International Congress Series*, Vol. 1282. Elsevier, 1026–1030.

[64] 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.

[65] 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.

[66] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1156–1160.

[67] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. 2019. Bbeep: A sonic collision avoidance system for blind travellers and nearby pedestrians. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[68] 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.

[69] Adarsh Kowdle, Yao-Jen Chang, Andrew Gallagher, and Tsuhan Chen. 2011. Active learning for piecewise planar 3d reconstruction. In *CVPR 2011*. IEEE, 929–936.

[70] 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.

[71] 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.

[72] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 151–162.

[73] Sooyeon Lee, Madison Reddie, Krish Gurdasani, Xiying Wang, Jordan Beck, Mary Beth Rosson, and John M. Carroll. 2018. Conversations for Vision: Remote Sighted Assistants Helping People with Visual Impairments. arXiv:1812.00148 [cs.HC]

[74] 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.

[75] 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).

[76] 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.

[77] Xiyan Liu, Gaofeng Meng, and Chunhong Pan. 2019. Scene text detection and recognition with advances in deep learning: a survey. *International Journal on Document Analysis and Recognition (IJDAR)* 22, 2 (2019), 143–162.

[78] Yang Liu, Noelle RB Stiles, and Markus Meister. 2018. Augmented reality powers a cognitive assistant for the blind. *ELife* 7 (2018), e37841.

[79] Srikanth Malla, Behzad Dariush, and Chiho Choi. 2020. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition. 11186–11196.

[80] 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.

[81] 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.

[82] Microsoft. 2021. Seeing AI - Talking camera app for those with a visual impairment. https://www.microsoft.com/en-us/ai/seeing-ai.

[83] Akihiro Miyata, Kazuki Okugawa, Yuki Yamato, Tadashi Maeda, Yusaku Murayama, Megumi Aibara, Masakazu Furuichi, and Yuko Murayama. 2021. A Crowdsourcing Platform for Constructing Accessibility Maps Supporting Multiple Participation Modes. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[84] MG Mohanan and Ambuja Salgoankar. 2018. A survey of robotic motion planning in dynamic environments. *Robotics and Autonomous Systems* 100 (2018), 171–185.

[85] Tim Morris, Paul Blenkhorn, Luke Crossey, Quang Ngo, Martin Ross, David Werner, and Christina Wong. 2006. Clearspeech: A Display Reader for the Visually Handicapped. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14, 4 (2006), 492–500. https://doi.org/10.1109/TNSRE.2006.881538

[86] 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.

[87] Muzammal Naseer, Salman Khan, and Fatih Porikli. 2018. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE Access* 7 (2018), 1859–1887.

[88] 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.

[89] Gerald Oster and Yasunori Nishijima. 1963. Moiré patterns. *Scientific American* 208, 5 (1963), 54–63.

[90] 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.

[91] 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.

[92] 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.

[93] Giorgio Presti, Dragan Ahmetovic, Mattia Ducci, Cristian Bernareggi, Luca Ludovico, Adriano Baratè, Federico Avanzini, and Sergio Mascetti. 2019. WatchOut: Obstacle sonification for people with visual impairment or blindness. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 402–413.

[94] 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.

[95] 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].

[96] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[97] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.

[98] 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

[99] 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.

[100] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. 2020. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research* 39, 8 (2020), 895–935.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence* 41, 9 (2018), 2035–2048.

[106] Sudipta N Sinha, Drew Steedly, Richard Szeliski, Maneesh Agrawala, and Marc Pollefeys. 2008. Interactive 3D architectural modeling from unordered photo collections. *ACM Transactions on Graphics (TOG)* 27, 5 (2008), 1–10.

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

[108] Jin Sun and David W Jacobs. 2017. Seeing what is not there: Learning context to determine where objects are missing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5716–5724.

[109] TapTapSee. 2021. TapTapSee. https://taptapseeapp.com/.

[110] 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.

[111] Ender Tekin, James M Coughlan, and Huiying Shen. 2011. Real-time detection and reading of LED/LCD displays for visually impaired persons. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 491–496.

[112] Nelson Daniel Troncoso Aldas, Sooyeon Lee, Chonghan Lee, Mary Beth Rosson, John M Carroll, and Vijaykrishnan Narayanan. 2020. AIGuide: An Augmented Reality Hand Guidance Application for People with Visual Impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13.

[113] 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.

[114] 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.

[115] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[116] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. 2021. From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network. *arXiv preprint arXiv:2108.09661* (2021).

[117] Zhihao Wang, Jian Chen, and Steven CH Hoi. 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[118] Galen Weld, Esther Jang, Anthony Li, Aileen Zeng, Kurtis Heimerl, and Jon E. Froehlich. 2019. Deep Learning for Automatically Detecting Sidewalk Accessibility Problems Using Streetscape Imagery. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 196–209. https://doi.org/10.1145/3308561.3353798

[119] Michele A Williams, Amy Hurst, and Shaun K Kane. 2013. "Pray before you step out" describing personal and situational blind navigation behaviors. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.

[120] Jingyi Xie, Madison Reddie, Sooyeon Lee, Syed Billah, Zihan Zhou, Chun-hua Tsai, and John Carroll. 2022. Iterative Design and Prototyping of Computer Vision Mediated Remote Sighted Assistance. *ACM Transactions on Computer-Human Interaction (in press)* (2022).

[121] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. 2018. Future person localization in first-person videos. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition. 7593–7602.

[122] Ruijie Yan, Liangrui Peng, Shanyu Xiao, and Gang Yao. 2021. Primitive Representation Learning for Scene Text Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 284–293.

[123] 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.

[124] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems* 30, 11 (2019), 3212–3232.

[125] 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.

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