Mobile Recognition and Tracking of Objects in the Environment through Augmented Reality and 3D Audio Cues for People with Visual Impairments

Oliver Beren Kaul
Kersten Behrens
Michael Rohs
<lastname>@hci.uni-hannover.de
Leibniz University Hannover
Hannover, Lower Saxony

ABSTRACT

People with visual impairments face challenges in scene and object recognition, especially in unknown environments. We combined the mobile scene detection framework Apple ARKit with MobileNet-v2 and 3D spatial audio to provide an auditory scene description to people with visual impairments. The combination of ARKit and MobileNet allows keeping recognized objects in the scene even if the user turns away from the object. An object can thus serve as an auditory landmark. With a search function, the system can even guide the user to a particular item. The system also provides spatial audio warnings for nearby objects and walls to avoid collisions. We evaluated the implemented app in a preliminary user study. The results show that users can find items without visual feedback using the proposed application. The study also reveals that the range of local object detection through MobileNet-v2 was insufficient, which we aim to overcome using more accurate object detection frameworks in future work (YOLOv5x).

CCS CONCEPTS

• **Human-centered computing** → *Accessibility theory, concepts and paradigms*; **Accessibility technologies**.

KEYWORDS

Mobile Scene Recognition, Accessibility, Visually Impaired, Spatial Audio, Text-to-Speech, Auditory Scene Description, Object Detection, Collision Warnings

ACM Reference Format:
Oliver Beren Kaul, Kersten Behrens, and Michael Rohs. 2021. Mobile Recognition and Tracking of Objects in the Environment through Augmented Reality and 3D Audio Cues for People with Visual Impairments. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451611

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451611

1 INTRODUCTION

People with visual impairments face major challenges when exploring unknown spaces or searching for objects, even in their own homes. Historically, only a few tools for exploring unknown spaces were available such as a guide dog or a white cane. Only a low percentage of the *blind and visually impaired persons* (VIPs) even utilize these aids (less than 10 % white cane users [38, 39] and about 2 % guide dog users [12] in the USA). Reasons for this can be found in the stigma of using a white cane or a guide dog, which distinguishes VIPs from other people. There are also issues with adapting to white cane usage, safety concerns [13], and the high cost for a guide dog and the need to care for it [12].

A solution for the stigma issues are hidden guidance devices, e.g. vibrotactile belts [35] or smartphone apps that provide navigation instructions and describe the world around the user based on data from private or public map services (e.g. Blindsquare [23], NavCog3 [30], Blavigator [9], BlindNavi [5], ASSIST [26]), or localized simultaneous localization and mapping (SLAM) approaches (e.g., Fusco et al. [10]). In recent years, several interesting works emerged in mobile scene recognition, and subsequent auditory scene description for people with visual impairments [1, 6, 22, 25, 26, 28, 32, 34, 40]. These approaches can be categorized into *online* and *offline/local* concepts. Online approaches take a picture using a smartphone camera, send this picture to a server for evaluation, and then provide an auditory description of the scene to the VIP. Using a remote server for scene recognition has several major disadvantages:

- The latency due to the communication with the server and the on-server processing.
- When the audio description is received on the smartphone and played back, the user might already have moved, resulting in inaccurate positioning of the 3D audio descriptions.
- For the same reason, detected objects cannot accurately be placed and tracked in the scene. This means they are lost if they are not detected in subsequent frames and cannot be used as static landmarks.
- Possible privacy implications from sending pictures or a video stream to an external server.

Offline/local approaches do not have these disadvantages but usually have a lower object recognition accuracy as a trade-off for recognition speed.

Our main contribution lies in merging concepts from different existing approaches [1, 8, 16, 22, 26, 28, 32, 34] into one compelling

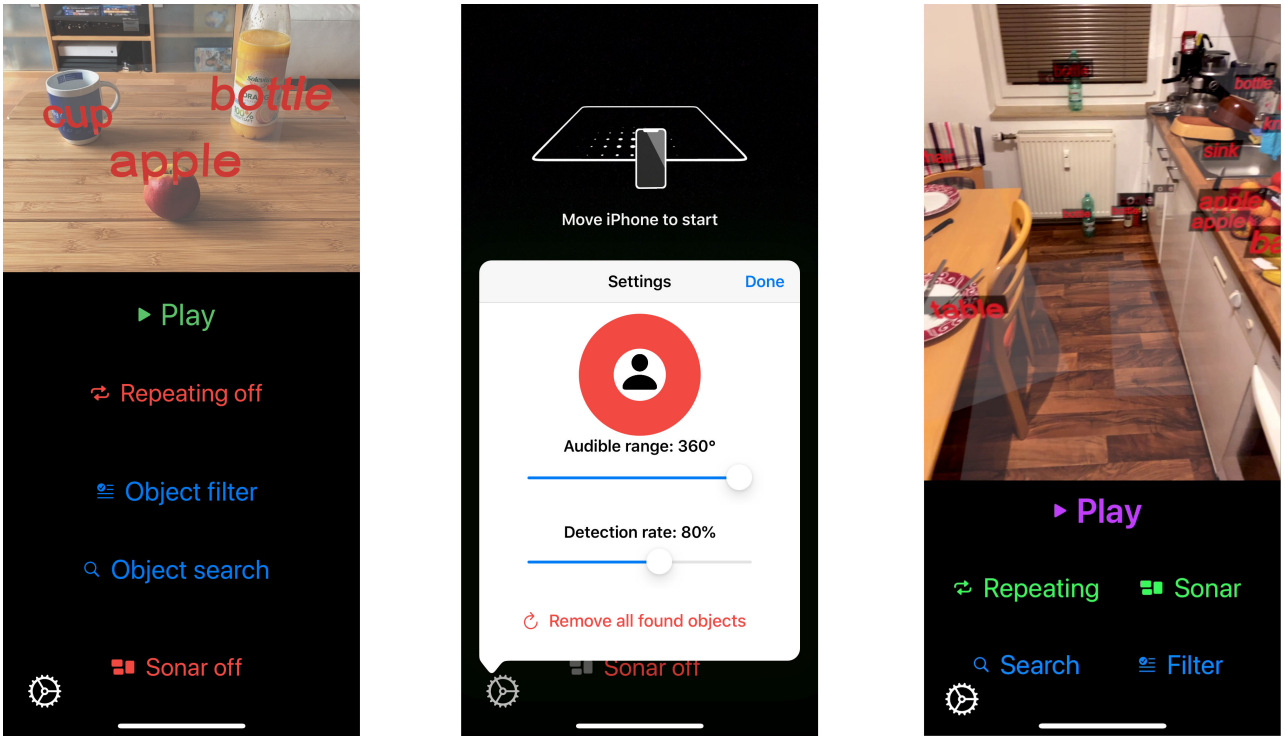


Figure 1: Left: the implemented system used in the study with MobileNet v2 as an object detection engine (range 1-1.5 m). Center: options for setting the auditory field between 0 and 360° and the detection threshold for the underlying neural network. Right: Preliminary future work version with YOLOv5x [11] (range 2-5 m).

iPhone app, which was validated in a preliminary usability study. Our app works as follows: First, we detect objects in camera images using MobileNet v2 [29] and in a future work version YOLOv5x [11]. These detected objects are placed in an ARKit scene at their respective depth. Placing objects into an AR scene has the major advantage of being able to use the objects as landmarks, even if the user turns away from the object or the object is no longer detected in subsequent frames. Additionally, it enables accurate 3D audio descriptions at their estimated locations, including their distance. Our app users can filter out or search for one or multiple specific object categories using iOS VoiceOver [3]. They may also tune the detection sensitivity and audible range of existing objects in the scene. Finally, a sonar-like sonification feature of nearby obstacles using ARKit's plane and wall detection helps users avoid running into walls and other obstacles.

2 RELATED WORK

Csapó et al. [7] give a good summary of developments up to 2014 of assistive technologies for the blind based on audio and tactile feedback.

A system consisting of glasses with a mounted camera, an Android phone for choosing desired objects, and a laptop to detect objects in pictures made by the camera is proposed by Thakoor et al. [33]. Feedback about object location, detected by a modified SURF algorithm, is relayed to the user by a 9-level auditory feedback method (e.g., left, up, center). They claim that their system is the first closed-loop system that provides object localization, recognition, and audio feedback for grabbing desired objects.

In terms of obstacle detection, Poggi et al. [27] propose a mobile system that detects objects through deep learning to give speech-based warnings of obstacles to VIPs. Using a tactile 3×3 grid on the abdomen, Van Erp et al. [36] present a system to indicate obstacle information around the user, including direction (3 levels), distance (4 levels), height (3 levels), and type (4 levels). They found that users had difficulties distinguishing the high amount of tactile patterns needed to identify the obstacle information. They found detection rates between 42 to 76 % for direction and height and 12.8 to 47 % for object distance after training. Van Erp et al. went for a multimodal pattern presentation approach (tactile and auditory) in their follow-up experiments [36].

Scene sonification is an exciting research direction, which allows VIPs to perceive a scene, including certain obstacles and navigation instructions, via auditory cues (through, e.g., flute or water drop sounds) [15]. Hu et al. [15] investigated three different scene sonification approaches (depth image sonification, obstacle sonification, and path sonification) in a comparative study. They found that preference for specific sonification approaches was highly individual. The sonification of high-level scene information (e.g., direction of a pathway) is generally easier to learn than low-level scene information (e.g., raw depth images).

A wearable assistive device aiming to navigate blind people in highly crowded urban areas is presented by Mocanu et al. [24]. They use a system consisting of a smartphone camera and ultrasonic

sensors. It can identify static and highly dynamic objects and warns the user about possible dangers using acoustic feedback. Li et al. [21] propose a system using an electronic SmartCane to assist VIPs with independent indoor travel. The system requires the location to be prepared by generating a semantic map. The system then detects dynamic and non-dynamic obstacles and provides the VIP with an adjusted path that avoids the obstacles on the way to the destination. Jośe et al. [19] provide a real-time assistance system that complements the white cane and is usable indoor and outdoor. The system can detect a path and obstacles within the path's borders and help avoid potential dangers. It guides the VIP around the obstacles while maintaining a walkable path.

Using the head-mounted AR platform Microsoft HoloLens, Eckert et al. [8] scan the user's surroundings. YOLOv2, a pre-trained neural network operating on a server back end, then analyzes the surroundings. The system gives feedback to the user in the form of directional 3D audio. The system could be extended to use the HoloLens depth information for obstacle avoidance. With *ReCog*, Ahmetovic et al. allow VIPs to train a neural network to detect their *personal items* at home and give auditory feedback about their locations [1].

A recent system similar to our approach is *WatchOut* by Presti et al. [28]. They developed an iOS app that uses ARKits' plane detection to sonify close objects to VIPs, similar to our sonar approach. However, Presti et al. did not include the type of object in the sonification and use it purely for obstacle avoidance. With *iVision*, Shen et al. [32] also developed an iOS app to search for single objects detected through YOLOv3 and sonified at their respective position using ARKit. This approach is very similar to ours but only allows searching for single objects (similar to our search function), does not place objects statically into the scene, and their study highlighted several usability issues with their app. *AIGuide* is another very recent system that allows users to search pre-identified ARKit objects in their home through sonification and tactile feedback. Unlike our approach, AIGuide can only identify a small set of pre-identified ARKit *ReferenceObjects*.

3 CONCEPT AND FIRST IMPLEMENTATION

Our first concept was informed by experiences from related work [1, 8, 16, 22, 26, 28, 32, 34]. The HoloLens application of Huang [16] showed that AR could increase the level of comfort and confidence of VIPs when doing search tasks by reading text present in the environment back to them. Additionally, Lin et al. [22] demonstrated an Android-based smartphone application that provides object detection, obstacle avoidance, and face detection. Our target population is VIPs with any degree of visual impairment from slightly impaired to fully blind. All VIPs can profit from an app helping them perceive objects and obstacles in the environment.

Initially, we investigated different mobile object detection frameworks and settled on using MobileNet v2 [29], as it could deliver an acceptable frame rate and decent recognition performance in our initial tests. YOLOv4 [4] and YOLOv5 [11] were not yet released. In terms of AR framework, we settled on using Apple ARKit [2], as many blind individuals own iPhones due to (initial) advantages [17] of iOS VoiceOver [3, 18] over Android Talkback [14, 18].

We decided to merge 3D audio descriptions of the local object detection provided by MobileNet v2 [29] with local AR scene recognition ARKit [2] as shown in Figure 2b, which provides the following benefits over existing scene description or scene sonification approaches:

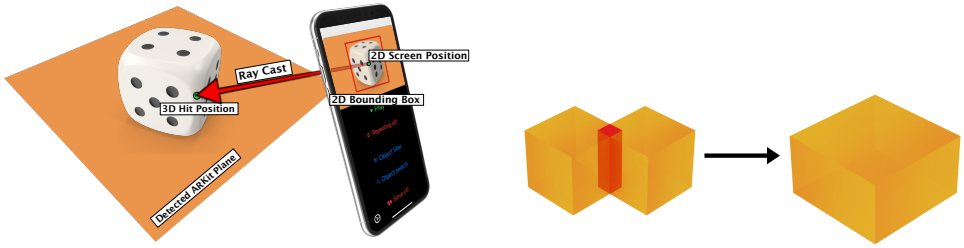
- Recognized items can be placed in the scene at a certain depth detected by the AR framework. Thus, it is possible to relay the distance of an item to the searching user, e.g., by using different volumes to different object descriptions or by 3D auditory distance descriptions in which the depth is described in natural language ("book at 2.5 meters, knife at 1 meter, ...").
- Recognized items can stay in the scene even if they are no longer detected in subsequent frames. This has the advantage that items can serve as landmarks, and users may be able to create a mental map of their surroundings more easily with the knowledge of certain items that are on the side, above, below, or behind them.
- Like other approaches, which detect the depth of obstacles using a depth camera [15], our approach can also detect obstacles and their position using ARKits plane detection. However, the current version of ARKit has some reliability issues, which will likely be fixed in future versions. Additional accuracy in plane and obstacle detection will be available through the depth camera present in the iPhone 12 Pro and future smartphones.

We implemented our concept as an iPhone app (system overview shown in Figure 3). Our approach includes obstacle detection and sonification through ARKit's [2] plane detection, and sonification through sonar-like sounds as well as 3D audio descriptions of detected objects by MobileNet v2 [29].

Both the sonar and the audio descriptions use a generic head-related transfer function (g-HRTF) [20] to provide 3D audio, including depth cues. While g-HRTFs initially perform worse in localization accuracy compared to personalized-HRTFs. They produce comparable results after some training and do not require individual calibration [20].

In terms of app options (see Fig. 1), the user may choose to only listen to obstacles and objects at a certain angle in front of her or him by adjusting the auditory field between 0 and 360°. The user may even fine-tune the neural network's detection threshold to either detect more objects with lower accuracy or fewer objects with more accuracy. Furthermore, we implemented an object filter to only read selected objects back to the user. We also implemented an object search function to find individual objects. All of these options are selectable using iOS VoiceOver [3].

A typical interaction with the system could look as follows: A VIP is searching for her handbag, which (unknown to her) was carried to a different place by her cat. 1. She starts the app. 2. She enters the search function using VoiceOver and selects "handbag". 3. She starts moving around her home, scanning the environment. 4. Entering her living room, the app detects the handbag laying on the floor to her right. The app starts vocalizing "handbag" to the lower right. 5. She repeatedly hears the direction and distance to the handbag and successfully picks it up.



(a) Determining distance and world position of an object by casting a virtual ray perpendicular to the screen from the center of its bounding box to the object with an underlying plane detected by ARKit in the real world.

(b) If the 3D bounds of an object overlap with an already existing object of the same category, both are merged into a single object to keep the scene tidy.

Figure 2: Merging ARKit plane detection and object detection by MobileNet v2 [29].

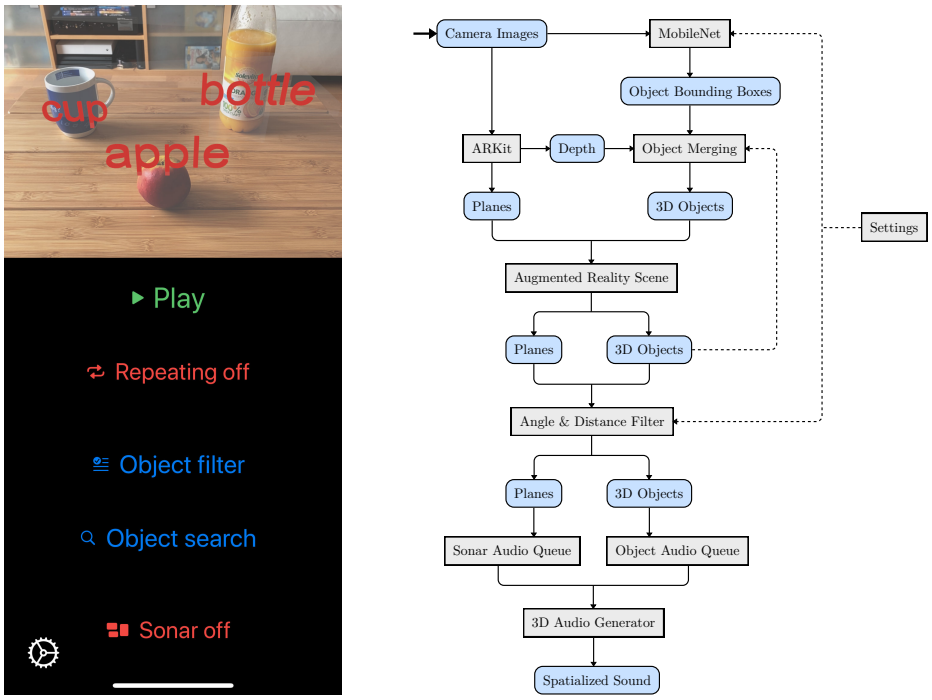


Figure 3: Implemented app (left) and abstract system overview (right).

4 EXPLORATORY USER STUDY

Using our first prototype, we decided to run an exploratory user study to gather qualitative feedback from blindfolded students and VIPs. We already knew from preliminary testing that the MobileNet v2 [29] object recognition accuracy and range would not be satisfactory. Thus, this study's focus is not on a search performance comparison vs. a baseline (e.g., white cane searching) but on qualitative feedback and usability of the app instead. We used the think-aloud method [37] and conducted the study in a lab room with five possible items as potential search targets (see Figure 4, left).

4.1 Procedure

On arrival, each participant had to fill out or agree to an informed consent form, which we read back to the blind participant and recorded his consent on audio. We further collected general participant data, including the kind and degree of visual impairment, through an introductory questionnaire. Subsequently, we introduced the participants to the iPhone app running on an iPhone 11 Pro and let them explore and try the options (shown in Figure 1) via iOS VoiceOver. In this **first experiment phase**, they could freely roam around the room and listen to audio feedback based on their settings while voicing concerns and feedback about the app to us.

In the **second experiment phase**, we fit the prototype belt with the iPhone 11 Pro and the app in the study mode on the participant's chest (see Figure 4, center). The settings for the study mode were detection rate 0.8 and auditory field 180°. These settings were determined in a pre-experiment by the authors. The participant's task was to find five potential targets, one after the other, by walking around, hearing target positions, and touching the targets. The

potential targets were a handbag, a teddy bear, a cup, a bottle, and a keyboard. The item positions were randomized between participants and between the two runs that each participant had to do. Our participants had to do two runs of the course, one with the app and, as a counterbalanced control condition, one with either their white cane or a makeshift white cane that we provided to participants who did not have their own. We measured search times in both conditions.

4.2 Participants

We invited a total of five participants (students) with normal vision who were blindfolded for the test (P1-P5) as well as one participant with visual impairments (P6, fully blind) for the study (all male, mean age 25 years, SD=5.5 years).

Another two participants with visual impairments (P7, male, 68 years old, fully blind, and P8, male, 70 years old, 30 % residual vision) explored the app and gave their oral feedback in the first study phase. We also consider their feedback in this study's qualitative results even though they did not participate in the second phase of the experiment due to time constraints and for safety reasons. We did not want to risk these two older participants running into tables or falling as the ARKit's plane detection [2] did not work perfectly, so our sonar feedback for obstacles was inaccurate sometimes. Figure 4 shows the younger blind study participant (35 years old) conducting the study with the smartphone app on the left and his white cane on the right.

4.3 Results and Discussion

In terms of quantitative search times of our participants in the second phase of the experiment, searching the five targets with the white cane took them on average 4m 53s (SD = 1m 3s), and searching with the app took them 5m 7s (SD = 2m 23s). P6 (blind) needed 4m 35s searching with his white cane and 2m 56s with the app, but he did the trials with his white cane first and created a mental map of the room in his head while doing the first run with his white cane. Keep in mind that this quantitative data is based on six participants, of which five were not trained on using a white cane or navigating blindly. Thus, the quantitative data is merely informational but cannot lead to firm conclusions.

Figure 5 shows the results of our final questionnaire on our app's usability. The results were mostly positive: all participants (slightly) agreed that they like the app. However, they saw the app more as a support for a white cane than a replacement. A reason could be that most participants could not imagine using the app outdoors. However, indoor navigation is a possible field of application, as all attendees could (fully) imagine using the app indoors. The application's interface was given general approval. All participants were able to navigate through the app safely. The feedback to the spatial audio features was very similar: the 3D audio was received very positively. Only one participant found the repetitive reading of the objects slightly disturbing, and all participants saw it as a helpful feature to read the objects. Regarding the sonar sound, the participants were more critical: Although nobody was disturbed by the sound, it was only helpful for three out of six (changing pitch), and four out of six participants did not feel insecure with the use of the sonar, respectively.

Apart from the general feedback collected by the questionnaire, we collected the following additional feedback by noting down vocal participant concerns while interacting with the app in both phases of the study. With special attention to the feedback from our blind participants, they noted that the app was an interesting prototype and that they could imagine using it to search for lost objects, e.g., in their own home but that the current detection rates and especially the range were too low, so they would likely still prefer to search everything by hand. The feedback from P8 was especially helpful, as he explored the app with his 30 % residual vision and made us realize the importance of strong contrasts within the app's buttons and detected objects apart from generally large fonts for VIPs who are not entirely blind.

We improved the app-based mostly on the feedback of our blind participants by increasing the contrasts of the detected objects, by applying a dark background, by enlarging the camera view while keeping the buttons rather large, and by switching from MobileNet v2 [29] to YOLOv5x [11]. YOLOv5x was released after conducting the study (see Figure 1, right).

5 LIMITATIONS

Our user study had several significant limitations, already hinted at earlier. First, blindfolded individuals have different experiences and abilities, and therefore they are not representative of the target population of VIPs [31]. Consequently, a future work study should solely include VIPs as participants. Secondly, a considerable limitation of our prototype in the user study was the relatively low detection accuracy of MobileNet v2 [29]. Several objects were only detected if the user was in close proximity of around 80 cm, while others were already detected at a distance of 1.5 m. Providing more useful information to the user without forcing him or her having to move all around the room would require a detection range above 3 m indoors to cover entire rooms (e.g., kitchens). This limitation can be solved by the recently released YOLOv4 [4], or YOLOv5 [11] frameworks, which significantly increase detection range and accuracy for everyday household objects to around 2-5 m while keeping an acceptable framerate of around 10 fps on an iPhone 11 Pro (YOLOv5x). We already have a prototype implementation of this concept (see Figure 1). However, we could not evaluate it in a study yet due to the COVID-19 limitations on conducting studies imposed by our university. Another issue of the implemented prototype was that a bug in ARKit was blocking the microphone for voice input (Siri). This glitch made it impossible to control the app via voice commands. Voice control was a highly desired feature among our blind participants and will likely be fixed in future versions of ARKit.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented the concept of merging local on-device object detection with an AR framework. Our first prototype implementation was still rather rough due to the low detection accuracy and the limited range of MobileNet v2 [29] and somewhat inexact plane detection of ARKit. Future work will address these issues by exchanging MobileNet v2 with YOLOv5x [11] and switching to an iPhone 12 Pro, which contains a depth camera that should significantly improve plane detection in ARKit. We plan to conduct

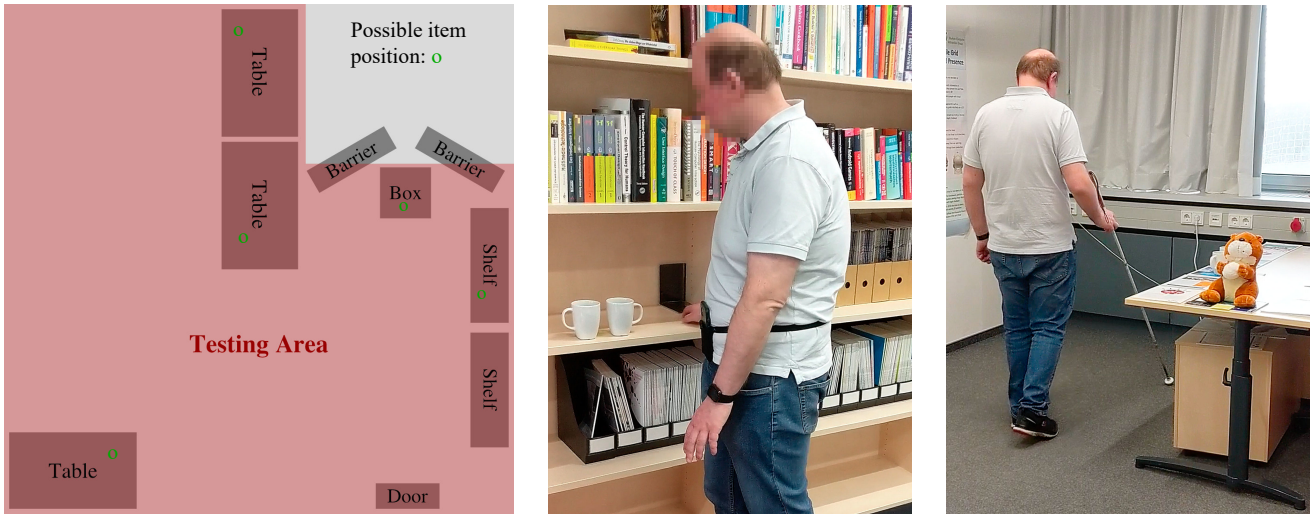


Figure 4: Left: Overview of our study room. The five possible item locations are marked with green circles. Center: Blind participant in the study, exploration of the study room with the app. Right: Blind participant in the study, exploration of the study room with his white cane.

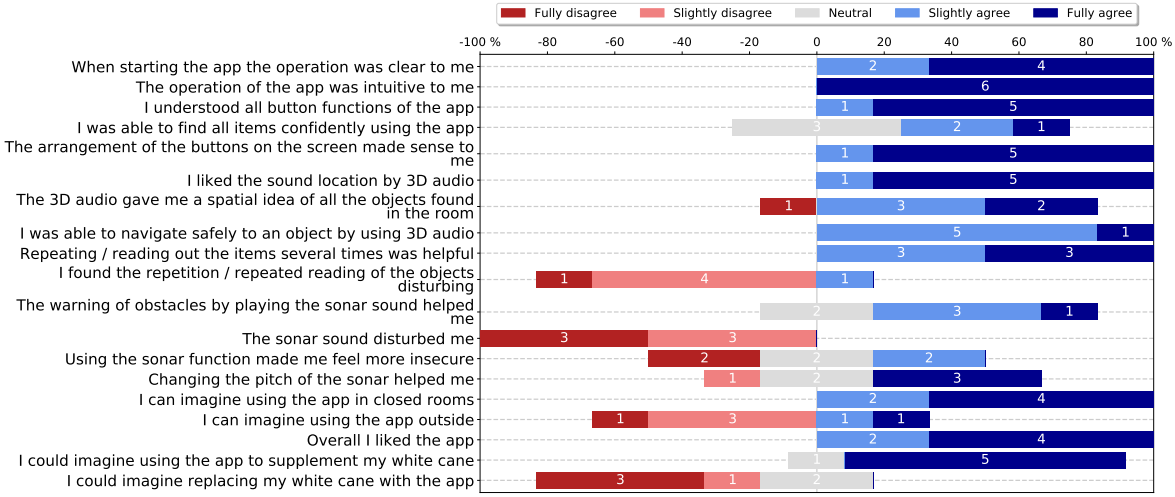


Figure 5: Subjective views of our study participants on the usability of the app.

a more extensive study with this much-improved prototype, which we already implemented but could not evaluate yet (see Figure 1, right). We expect it to be precious to VIPs from our preliminary tests with the improved prototype, as it does not require any additional hardware beyond an iPhone and headphones. It runs entirely locally on the users' phones, protecting their privacy. It allows the user to simply hear the type, direction, and distance of objects and obstacles in their surroundings in real-time, without the limitations caused by an online approach (e.g., variable latency and connectivity issues). Another exciting feature, which remains to be validated in a user study, is the possibility to use the detected objects as landmarks in the scene, possibly even listening to their position when they are behind the user and no longer recorded on video.

REFERENCES

[1] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. 2020. ReCog: Supporting Blind People in Recognizing Personal Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376143

[2] Apple Inc. 2020. Apple ARKit. https://developer.apple.com/augmented-reality/arkit/

[3] Apple Inc. 2021. Apple Accessibility - VoiceOver. https://www.apple.com/accessibility/vision/

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 http://arxiv.org/abs/2004.10934

[5] Hsuan-Eng Chen, Yi-Ying Lin, Chien-Hsing Chen, and I-Fang Wang. 2015. BlindNavi: A Navigation App for the Visually Impaired Smartphone User. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, Vol. 18. ACM, New York, NY, USA, 19–24. https://doi.org/10.1145/2702613.2726953

[6] CloudSight Inc. 2020. TapTapSee App. https://taptapseeapp.com/

[7] Ádám Csapó, György Wersényi, Hunor Nagy, and Tony Stockman. 2015. A survey of assistive technologies and applications for blind users on mobile platforms: a review and foundation for research. *Journal on Multimodal User Interfaces* 9, 4 (dec 2015), 275–286. https://doi.org/10.1007/s12193-015-0182-7

[8] Martin Eckert, Matthias Blex, and Christoph M Friedrich. 2018. Object Detection Featuring 3D Audio Localization for Microsoft HoloLens - A Deep Learning

based Sensor Substitution Approach for the Blind. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, Vol. 5. SCITEPRESS - Science and Technology Publications, -, 555–561. https://doi.org/10.5220/0006655605550561

[9] Hugo Fernandes, Paulo Costa, Hugo Paredes, Vítor Filipe, and João Barroso. 2014. Integrating Computer Vision Object Recognition with Location Based Services for the Blind. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8515 LNCS. Springer Verlag, -, 493–500. https://doi.org/10.1007/978-3-319-07446-7_48

[10] 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*. ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3371300.3383345

[11] Glenn Jocher. 2020. YOLOv5 on Github. https://github.com/ultralytics/yolov5

[12] Guiding Eyes for the Blind. 2020. How many people use guide dogs? https://www.guidingeyes.org/about/faqs/

[13] Marion Hersh. 2015. Cane use and late onset visual impairment. *Technology and Disability* 27, 3 (2015), 103–116. https://doi.org/10.3233/TAD-150432

[14] Jerry Hildenbrand. 2014. What is Google TalkBack? https://www.androidcentral.com/what-google-talk-back

[15] Weijian Hu, Kaiwei Wang, Kailun Yang, Ruiqi Cheng, Yaozu Ye, Lei Sun, and Zhijie Xu. 2020. A comparative study in real-time scene sonification for visually impaired people. *Sensors (Switzerland)* 20, 11 (2020), 1–17. https://doi.org/10.3390/s20113222

[16] Jonathan Huang, Emily Cooper, and Wojciech Jarosz. 2017. *A HoloLens Application to Aid People who are Visually Impaired in Navigation Tasks.* Technical Report. Dartmouth College.

[17] Incobs. 2014. A comparison of speech output of VoiceOver, TalkBack and Narrator | INCOBS. https://www.incobs.de/tests-english/items/a-comparison-of-output-of-voiceover-talkback-and-narrator.html

[18] Danielle Irvine, Alex Zemke, Gregg Pusateri, Leah Gerlach, Rob Chun, and Walter M Jay. 2014. Tablet and Smartphone Accessibility Features in the Low Vision Rehabilitation. *Neuro-Ophthalmology* 38, 2 (2014), 53–59. https://doi.org/10.3109/01658107.2013.874448

[19] J. Jośe, J. M.H. Du Buf, and J. M.F. Rodrigues. 2012. VISUAL NAVIGATION FOR THE BLIND - Path and Obstacle Detection. In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, Vol. 2. SciTePress - Science and and Technology Publications, -, 515–519. https://doi.org/10.5220/0003711405150519

[20] Florian Klein and Stephan Werner. 2016. Auditory Adaptation to Non-Individual HRTF Cues in Binaural Audio Reproduction. *Journal of the Audio Engineering Society* 64, 1/2 (feb 2016), 45–54. https://doi.org/10.17743/jaes.2015.0092

[21] Bing Li, Juan Pablo Munoz, Xuejian Rong, Qingtian Chen, Jizhong Xiao, Yingli Tian, Aries Arditi, and Mohammed Yousuf. 2019. Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People. *IEEE Transactions on Mobile Computing* 18, 3 (mar 2019), 702–714. https://doi.org/10.1109/TMC.2018.2842751

[22] Bor-Shing Lin, Cheng-Che Lee, and Pei-Ying Chiang. 2017. Simple Smartphone-Based Guiding System for Visually Impaired People. *Sensors* 17, 6 (jun 2017), 1371. https://doi.org/10.3390/s17061371

[23] MIPsoft. 2020. Blindsquare App. https://www.blindsquare.com/de/

[24] Bogdan Mocanu, Ruxandra Tapu, and Titus Zaharia. 2016. When Ultrasonic Sensors and Computer Vision Join Forces for Efficient Obstacle Detection and Recognition. *Sensors* 16, 11 (oct 2016), 1807. https://doi.org/10.3390/s16111807

[25] Vikky Mohane and Chetan Gode. 2016. Object recognition for blind people using portable camera. In *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*. IEEE, -, 1–4. https://doi.org/10.1109/STARTUP.2016.7583932

[26] Vishnu Nair, Greg Olmschenk, William H Seiple, and Zhigang Zhu. 2020. ASSIST: Evaluating the usability and performance of an indoor navigation assistant for blind and visually impaired people. *Assistive Technology* 0, 0 (sep 2020), 1–11. https://doi.org/10.1080/10400435.2020.1809553

[27] Matteo Poggi and Stefano Mattoccia. 2016. A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning. *Proceedings - IEEE Symposium on Computers and Communications* 2016-August (2016), 208–213. https://doi.org/10.1109/ISCC.2016.7543741

[28] 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*. ACM, New York, NY, USA, 402–413. https://doi.org/10.1145/3308561.3353779

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 0, 0 (jan 2018), 4510–4520. arXiv:1801.04381 http://arxiv.org/abs/1801.04381

[30] Daisuke Sato, Uran Oh, Kakuya Naito, Hironobu Takagi, Kris Kitani, and Chieko Asakawa. 2017. NavCog3: An evaluation of a smartphone-based blindindoor navigation assistant with semantic features in a large-scale environment. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 270–279. https://doi.org/10.1145/3132525.3132535

[31] Andrew Sears and Vicki L Hanson. 2012. Representing users in accessibility research. *ACM Transactions on Accessible Computing* 4, 2 (mar 2012), 1–6. https://doi.org/10.1145/2141943.2141945

[32] Jinyang Shen, Zhanxun Dong, Difu Qin, Jingyu Lin, and Yahong Li. 2020. iVision: An Assistive System for the Blind Based on Augmented Reality and Machine Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12188 LNCS. Springer, -, 393–403. https://doi.org/10.1007/978-3-030-49282-3_28

[33] Kaveri Thakoor, Nii Mante, Carey Zhang, Christian Siagian, James Weiland, Laurent Itti, and Gérard Medioni. 2015. A System for Assisting the Visually Impaired in Localization and Grasp of Desired Objects. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8927. Springer Verlag, -, 643–657. https://doi.org/10.1007/978-3-319-16199-0_45

[34] 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*, Vol. 13. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3373625.3417028

[35] Koji Tsukada and Michiaki Yasumura. 2004. ActiveBelt: Belt-Type Wearable Tactile Display for Directional Navigation. In *Ubiquitous Computing*. Vol. 3205. Springer, Berlin, Heidelberg, 384–399. https://doi.org/10.1007/978-3-540-30119-6_23

[36] Jan B.F. van Erp, Liselotte C.M. Kroon, Tina Mioch, and Katja I. Paul. 2017. Obstacle detection display for visually impaired: Coding of direction, distance, and height on a vibrotactile waist band. *Frontiers in ICT* 4, SEP (2017), 1–19. https://doi.org/10.3389/fict.2017.00023

[37] M W Van Someren, Y F Barnard, and J A C Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive processes.

[38] Washington State. 2020. Dispelling Myths, Department of services for the blind. https://dsb.wa.gov/resources/blind-awareness/dispelling-myths

[39] White Cane Day. 2020. White Cane Day FAQ. http://whitecaneday.org/canes/

[40] Yu Zhong, Pierre J. Garrigues, and Jeffrey P. Bigham. 2013. Real time object scanning using a mobile phone and cloud-based visual search engine. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/2513383.2513443