Audi-Exchange: AI-Guided Hand-based Actions to Assist Human-Human Interactions for the Blind and the Visually Impaired

Daohan Lu[1,2] and Yi Fang*[1,2,3]

[1]NYU Multimedia and Visual Computing Lab, Abu Dhabi and New York
[2]New York University, New York, NY 10003, USA
[3]New York University Abu Dhabi, Abu Dhabi 129188, UAE
Email: {dl3957,yfang}@nyu.edu

Abstract

Vision loss or low vision poses significant challenges to blind-or-visually-impaired (BVI) individuals when interacting with humans and objects. Although many apps and assistive devices can help them better interact with the environment and objects, the current state of assistive technology leaves human-human interaction needs of the BVI largely unaddressed. Because of this, we introduce a new wearable mobile assistive platform, named Audi-Exchange, to address part of the problem. Developed with mobile-optimized computer vision and audio engineering techniques, Audi-Exchange facilitates a specific area of human-human interaction by helping the BVI user accurately locate another person's hand with spatial audio in order to pass objects over to or receive objects from the other person. Audi-Exchange differs from existing academic and commercial assistive technologies in that it is intuitive to use and non-intrusive when worn. We conduct several experiments to investigate Audi-Exchange's effectiveness as an assistive human-human interaction tool and discover encouraging results.

1. Introduction

In the workplace or daily life, people need to interact with other people for various social and working needs. However, people who are blind or visually impaired (BVI) have fewer or lesser tools available to them for interacting with another person, namely those that rely on a clear vision, such as sensing eye contact or reading body language. This leads to a harder time when they interact with other people, especially with those who are not familiar with the appropriate practices for interacting with BVI individuals. This can lead to strong real-world impacts. In the workplace, for instance, visual impairment leads to lower efficiency when interacting and communicating with coworkers and customers, which makes employment more difficult. Thus, we wish to investigate the feasibility and effectiveness of developing an assistive device to collect important visual cues in the environment and relays this information to the BVI user to help the BVI user interact with other sighted or BVI people. In this paper, we focus on facilitating human-human interaction that involves handing over of objects. The choice is due to the common situation in which a BVI person needs to exchange objects with another person in casual settings as well as in the workplace. For instance, a BVI person may need to accept a credit card from someone or offer a cup of water to someone. To help the BVI with these types of actions, we introduce the Audi-Exchange wearable mobile assistive platform. When Audi-Exchange is activated, it collects images from a camera and uses an efficient convolution neural network (CNN) to detect another person's hand within the camera's field of view. When a hand is detected, Audi-Exchange tracks its on-camera location and computes its corresponding 3D direction based on the camera's optical parameters. Next, the direction of the target hand is relayed to the BVI user through stereo headphones as an audio tone, which is processed by a head-related transfer function (HRTF) to appear as coming from the computed direction. In this fashion, we leverage computer vision and audio engineering to augment or substitute impaired human vision to allow the BVI to "hear" how they should move when handing over or receiving an object from another person. We assess Audi-Exchange's effectiveness and speed by evaluating it on a proof-of-concept system consisting of a personal computer and a camera through several experiments. Lastly, we show that the system could be ported to work with mainstream smartphone hardware with a few modifications to the algorithm and discuss the next step for Audi-Exchange.

2. Assistive Technology Landscape

When reviewing academic sources, we found an abundance of research that aimed to develop assistive devices for the blind and visually impaired. A large portion falls into the category of electronic travel aids (ETAs), which are devices that gather information, including nearby objects and obstacles, about the surrounding environment via dedicated sensors and transfer it to the user [12, 21]. Some examples include an RFID-based indoor navigation system [16], the smart cane [24], the Path Force Feedback Belt [19], and Substitute Eyes (a hand-worn ultrasonic obstacle warning device) [11]. Another category that assistive devices fall into is vision substitution, which is using computer vision to capture images of the surroundings and transforming the raw visual information into a VI-friendly form, such as haptics or sounds. Vision substitution devices are more general-purpose as the information gathered by the system is less condensed compared to ETAs and relayed in a more direct and visual form to the VI, with the exception being text recognition sometimes is performed on images to read text printed in the surroundings to the user. For instance, FingerReader reads printed text to the user with a hand-worn camera [22]. Silicon Eyes informs the user of the color of objects nearby in addition to facilitating outdoor navigation [20]. Overall, however, even though many assistive devices developed in the academic community exist that cater to different aspects of BVI assistance, most have significant drawbacks. Mainly, they can be cumbersome due to being equipped with complex sensors and can be intrusive due to needing to be worn at various locations around the body to effectively gather information. Also, few papers mention estimated production costs for the the assistive devices that were proposed, which leads to uncertainty as to whether these devices can be efficiently built and widely distributed.

There also exists a number of commercial assistive devices for the VI. A major portion are visual enhancement tools that are designed to be used by the low vision and not fully blind (perhaps because products made for the low vision enjoy a larger market compared to those made for the fully blind). The basic white cane is arguably the most widely used [10, 14] and affordable tool useful to both the low vision and fully blind, but its function is limited by its short range as a direct extension of physical touch. Advancement in digital imaging sensors led to a variety of vision enhancement tools that process images of the surroundings using camera sensors and image processing techniques (such as zooming in or boosting contrast). These include IrisVision [7], Acesight [5], and eSight [6]. Vision enhancement can also be accomplished with certain accessibility apps running on Google Glasses [1] and Microsoft Hololens [2]. More advanced solutions exist (such as the Orcam MyEye 2.0 [8]) where visual information gathered by the camera sensors is condensed into audio notifications for text reading, identifying objects, and recognizing faces. Even though commercial solutions are generally more comfortable to use, they are generally very costly. The vision augmentation devices mentioned above come in the price range of around $2,000-4,000 USD, with OrCam MyEye 2.0 [8] being the most expensive, costing $4,250 at the time of writing. Despite having being somewhat common, few hardware-based assistive solutions made by researchers or commercial firms are designed to specifically cover the VI's human-human interaction needs.

On the other hand, purely software-based solutions can be run on common smartphones and are more affordable, with Microsoft Seeing AI and BlindSquare [4, 9] being popular free vision enhancement and BVI navigation apps. However, software-based solutions generally also have a set of significant shortcomings as they tend to (1) lack on-board visual processing, which leads to functional reliance on online computing, or (2) like the hardware-based solutions discussed previously, they only facilitate interactions with the environment while the VI's human interaction needs are largely ignored. In summary, despite innovations in mobile information gathering and visual computing that enabled many innovative hardware and software assistive solutions for the BVI, there are still only few that are designed specifically to help the user have smoother, richer interactions with other people.

3. Motivation and Merit

Because existing assistive technology for the BVI is often costly to obtain, uncomfortable to use, and does not address human interaction needs (fig. 1), we wish to design a new assistive device to address these problems. A 2020 comprehensive literature review on wearable device design reveals that "comfort", "intuitiveness", and "mobility" are the three most desired properties of a wearable device in terms of user experience [13]. Comfort corresponds to the minimization of discomfort or pain, so the device should have a small form factor, be worn on convenient locations, and have good thermal dissipation [13]. For "intuitiveness", the human-computer interaction pattern should be easy to learn and it should resemble familiar interactive patterns used in everyday life. For "mobility", the device should be lightweight and not obstruct the user's physical activity. These three properties are not exclusive to health-sighted individuals and are arguably more important for blind or visually impaired individuals because they expect a higher degree of usability and reliability from assistive wearable devices. As a consequence, we design a new assistive device that specifically aims to facilitate human-human interactions while maximizing comfort, intuitiveness, and mo-

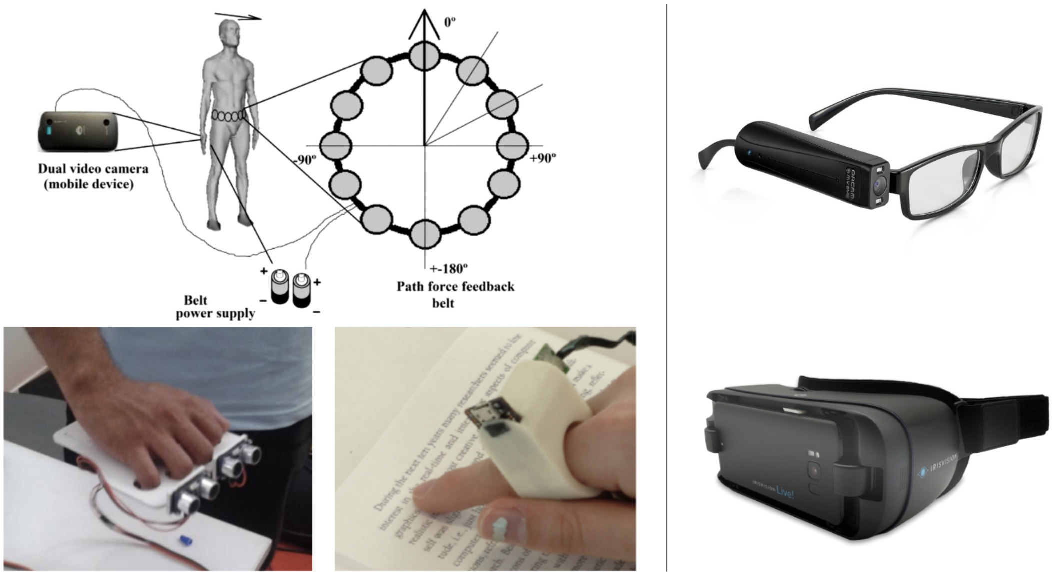

Figure 1. Existing assistive device solutions for the BVI can be cumbersome, intrusive (first column), or costly (second column). Also, few devices are designed to address human-human interaction needs. The three images in the first column read from left to right and then top to bottom are: the Path Force Feedback Belt [19], Substitute Eyes [11], and FingerReader [22]. The two images in the second column read from top to bottom are: OrCam MyEye 2 [8] and IrisVision Live! [7], which sell for $4,250 and $2,950 respectively at the time of writing.

bility when worn. We think that good user experience is critical to the actual rate of adoption in the BVI community and the social impact that results from it.

Named Audi-Exchange, we propose an assistive platform that uses real-time onboard mobile computer vision and audio engineering to help the BVI better handle hand-based human-human interactions. From the user's perspective, Audi-Exchange sends spatial audio cues through a pair of headphones to help the user locate another person's hand to hand over or receive an object. For example, it helps a blind cashier reach over the counter to receive a credit card from a customer. From a technical standpoint, Audi-Exchange uses a smartphone's camera and processor and mobile-optimized computer vision algorithms to track the hand motion of a person next to the BVI user and relay this information in real-time to the BVI user through the stereo headphones via spatial audio. In this manner, Audi-Exchange is designed to be responsive, energy-efficient, intuitive, comfortable, and mobile. Unlike assistive technologies that rely on dedicated sensors and processors, all parts of Audi-Exchange can run on mainstream smartphones using onboard mobile CPU/GPU/Neural Processor as the "brain", the camera as the sensory input, and a pair of wired or wireless headphones as the sensory output. Regular headphones be optionally replaced with bone conduction headphones for better perception of ambient sounds.

4. Method

4.1. Hand Tracking

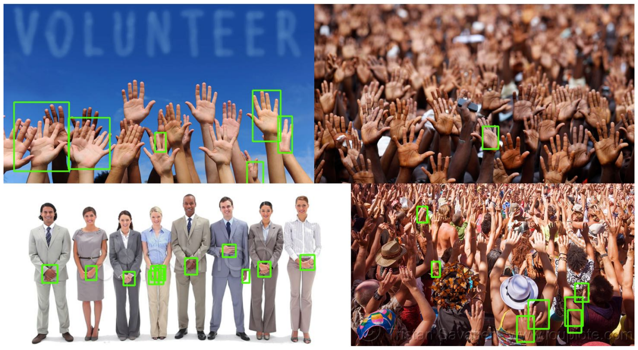

Figure 2. Some images depicting the performance of the hand tracking algorithm [23] we adopt. Image credit to [23].

To implement Audi-Exchange, it is necessary to have a fast and accurate hand-tracking algorithm for the computer program to determine the location of the interactive partner's hand on an image. Hence, we base the hand tracking functional unit of Audi-Exchange on an efficient hand tracking algorithm [23] utilizing MobileNet (a mobile-optimized computer vision architecture) [15] and SSD (a computationally efficient object detection method) [17]. We did not fine modify or fine tune the hand tracking model by [23] during our experiments. Running on a 2014 13" Macbook Pro with a 2.6GHz dual-core Intel Core i5, we found the hand tracking algorithm to be able to process an average of 13 frames per second (FPS) with an input image size of 320 by 180 pixels. Such a speed is sufficient for tracking a non-fast-moving hand in real-time. Note that modern smartphones are typically equipped processors that can outperform the 2014 Macbook Pro by significant margins and many even come with dedicated neural processors that further speed up computer vision and machine learning applications, as is the case with the latest Apple iPhone (dubbed the "Neural Engine") and Google Pixel (dubbed the "Neural Core"). As a result, the 13-FPS speed that we recorded represents a bottom-line performance figure when hand tracking is run on older, less capable devices. We show in fig. 2 some test images of the hand tracking results as published by the author [23], while our test results are detailed in sec. 5.1.

4.2. 3D Audio

After a hand is detected, we wish to transmit this information to the BVI user in the form of an audio signal. To this end, we develop the 3D Audio module, which utilizes the OpenAL API [3] to embed 3D location cues into an arbitrary sound source. The method in which OpenAL (and other similar APIs like Resonance Audio [18]) accomplishes this is through the head-related transfer function (HRTF). The HRTF is a transfer function that takes in an artificial source of sound, denoted as $x(f)$, which is a function of frequency, and outputs separate channels of audio to the left (denoted $X_L(f)$) and right (denoted $X_R(f)$) ears of the listener to create directional sound [25]. When given HRTFs $H_L(f,\theta,\phi)$ and $H_R(f,\theta,\phi)$ and the source audio $x(f)$, the output $X_L(f)$ and $X_R(f)$ can be computed as:

$$\begin{aligned} X_L(f) &= H_L(f,\theta,\phi) * x(f) \\ X_R(f) &= H_R(f,\theta,\phi) * x(f) \end{aligned} \tag{1}$$

where θ, ϕ are known constants representing the azimuth and elevation of the simulated audio source. Cartesian (x, y, z) coordinates can be used too, though only the radial direction will be taken into account by the most HRTFs with the distance determining volume. In this way, original audio signals are filtered to carry directional cues to the listener. We utilized this fact to help the user locate a virtual audio source whose location corresponds to the actual location of the hand being tracked. When given a 3D location, the 3D Audio module plays a tone processed by the HRTF so that it appears as coming from the given location when the user hears it through stereo headphones. By relaying visual information to the user through sound, 3D Audio serves as a crucial part that makes it possible for the device to communicate visual information to a BVI user. Based on our testing, we found the 3D Audio module to consume negligible computational resources when it processes and plays spatial audio, making it compatible with the mobile-optimized design of Audi-Exchange.

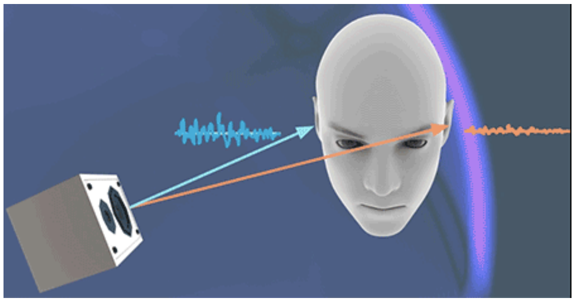

Figure 3. A graphic illustrating the Interaural time differences (ITD) used in 3D audio software to help humans determine the horizontal location of low-frequency sounds. Image source: Resonance Audio [18]

4.3. Audi-Exchange

Finally, we combine the hand tracking and 3D audio modules to build the Audi-Exchange prototype. It works as follows: in sequential order, a camera captures an image, which is sent to the hand tracking module. After the hand location is determined, a virtual sound source is created with 3D audio to reflect the horizontal location of the detected hand. As the user hears the 3D sound through headphones, they acknowledge the location of the partner's hand in the real world. Then, the user moves their own hand to the perceived location in an attempt to touch the other person's hand. The process repeats until the action is completed as determined by the user. The flow of information from the BVI user's perspective is illustrated in fig. 4. By offering auditory feedback in response to visual interactive cues (hand motion), Audi-Exchange repairs or augments hand-eye coordination for VBI user to efficiently hand objects over to or receive objects from another person. Although the Audi-Exchange test system used in this paper is set up for laptops and desktop computers, it is simple to port the program to mobile devices by migrating to open-source mobile-based computer vision and 3D audio algorithms (see sec. 6). A concept design of the mobile-based hardware is pictured in fig. 5.

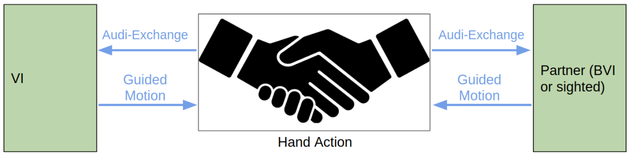

Figure 4. Audi Exchange relays information about the location of the interactive partner's hand to the BVI user to guide the user toward handing over to or receiving an object from another person, who may be VIB or sighted.

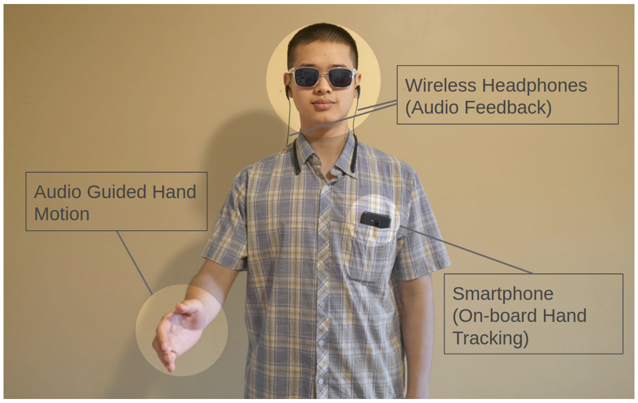

Figure 5. Audi-Exchange Concept Design. By tracking the interactive partner's hand on a smartphone and sending spatial-audio-based interactive cues via wireless headphones, Audi-Exchange helps the BVI with hand-related human-human interactions in a lightweight and comfortable package.

5. Experiments

5.1. Hand Tracking Experiment

In this experiment, we wish to find how reliably a mobile hand tracking program can pick up the interactive partner's hand. We designed four different test scenes to investigate some realistic settings under which Audi-Exchange would be used. They are categorized by the lighting conditions and the object that the partner's hand is holding. The categories are 1. Regular Lighting / Empty Hand, 2. Regular Lighting / Holding Pen, 3. Regular Lighting / Holding Cup, and 4. Poor Lighting / Empty Hand. We show one image from each test case in fig. 6. The first two test scenes happen most frequently, with the interactive partner shown on the camera trying to receive an object by extending an empty hand or trying to give a small object with their hand. The third test scene exemplifies situations in which the partner offers a large object that occludes part of their hand when seen from the camera's perspective. The fourth test scene aims to see how well the hand can be tracked when the partner's hand is not properly lit and light is coming from behind the partner. All test scenes are filmed indoors with warm-colored LED lighting. A smartphone camera is used captures a person facing the camera from head-height, which simulates the BVI user's egocentric view. The person in view moves the extended hand around throughout the video to create natural motion blur and location variations. The smartphone camera used has a horizontal field of view of approximately 90° and an aspect ratio of 16 by 9, with the shutter speed set to $1/50$ of a second, ISO to 1200, and white balance to "Auto". All images are resized to 320 by 180 pixels before being sent to the hand tracker.

To quantitatively evaluate tracking performance, we record the number of true positives, false positives, and false negatives for each test scene (see tab. 1). A true positive is defined as when the hand offering or receiving an object is detected or when both hands of the partner are detected (which is a trivial issue, as the hand we're interested in will be elevated and extended forward, meaning it can be distinguished from the other hand by picking the topmost hand). For similar reasons, exactly two closely overlapping bounding boxes on the correct hand is also considered a true positive, as the issue can be trivially solved by treating it as if two hands were detected. A false positive is defined as when the hand of interest is not detected, but an irrelevant object or the hand not of interest is. In the "False Positive (Wrong Hand)" column, we show the number of false positive frames in which the false positive is due to detecting only the irrelevant hand and no other objects. Lastly, a false negative is defined by not detecting any object or hand in a frame. Note that in all test cases, all frames include the hand of interest, so the true negative statistic is not available (or, equivalently, it has 0 instances). For each test case, there are 50 randomly selected frames from a video sequence. A frame has an equal chance of having the person and hand appear in the middle, left, and right parts of the image. A human annotator determines whether a frame is considered a true positive, false positive, false positive – wrong, or false negative by observing the visual output of the hand tracking algorithm. The annotator moves on the next frame when ready. The quantitative result is shown in tab. 1. Select qualitative hand tracking outputs are shown in fig. 7. We include a video segment in the supplementary materials demonstrating the hand tracking software used in this experiment.

	T.P.	**F.P.**	F.P. (Wrong Hand)	**F.N.**
Test 1	46	3	2	1
Test 2	47	2	0	1
Test 3	21	22	15	7
Test 4	0	50	0	0

Table 1. The recorded instances of true positives (T.P.), false positives (F.P.), part of false positives that are due to detecting the irrelevant hand (F.P. [Wrong Hand]), and false negatives (F.N.) for each hand tracking test scene. A total of 50 frames were examined for each test scene. Tests 1-4 correspond to Regular Lighting / Empty Hand, Regular Lighting / Holding Pen, Regular Lighting / Holding Cup, and Poor Lighting / Empty Hand, in this order.

We found the hand tracking algorithm to be robust under regular lighting conditions for tracking both the receiving and giving hand poses in the "Empty hand' and "Holding pen" test scenes. Even when a significant portion of the hand is occluded by an object, which is the case for the "Holding Cup" test scene, the algorithm was still able to detect the hand of interest 42% of the time. If the hand of

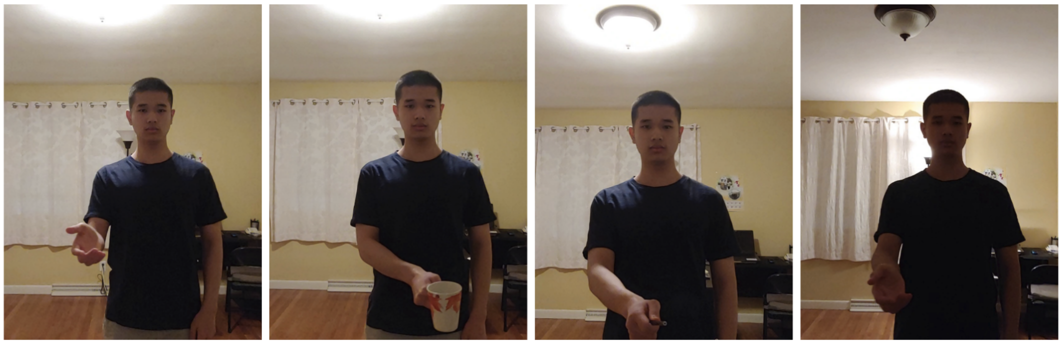

Figure 6. Overview of the four test cases. From left to right: 1. Regular Lighting / Empty Hand, 2. Regular Lighting / Holding Pen, 3. Regular Lighting / Holding Cup, and 4. Poor Lighting / Empty Hand. We choose frames where the subject and hand are in the center of the image for clarity. Zoom in for better details.

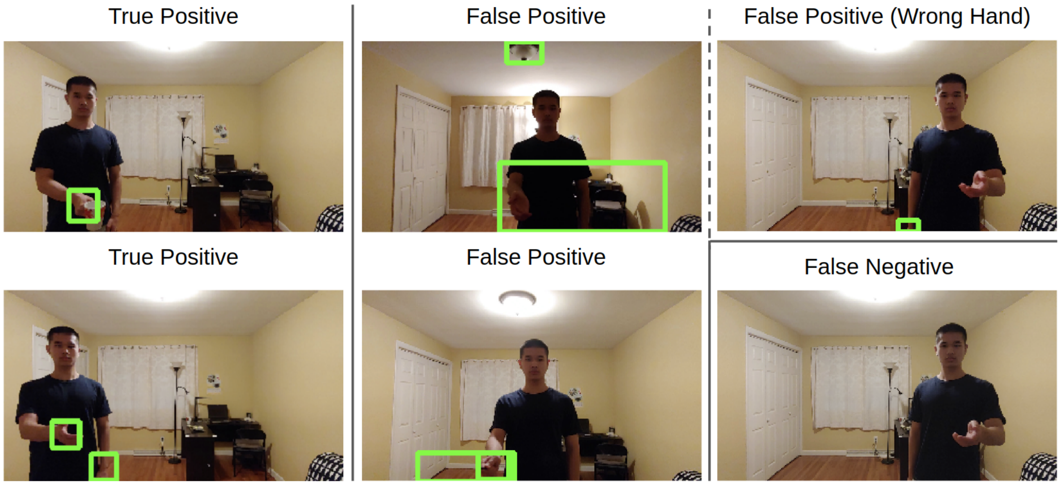

Figure 7. Some examples of the visual output of the hand tracking algorithm. Each is annotated as True Positive, False Positive, False Positive (Wrong Hand), or False Negative Classes. Note that a “False Positive (Wrong Hand)” image counts toward False Positive also. Zoom in for more details.

interest was not detected, the other hand is detected 51.7% of the time, which to some extent still reflects the location of the hand of interest as the two hands usually appear at nearby locations. In poor lighting, however, the algorithm fails to reliably track the hand. Specifically, the unlit overhead light in the room was picked up as a hand in all 50 frames tested. Another finding worth mentioning was that tracking performance was insensitive to the on-camera location of the hand. Hands appearing on the edges of a frame are tracked with approximately the same accuracy as those near the center. Based on our testing, we believe the mobile-optimized hand-tracking algorithm [23] selected for this experiment to be sufficiently reliable to locate hands in real-time when the hands are properly lit and exposed.

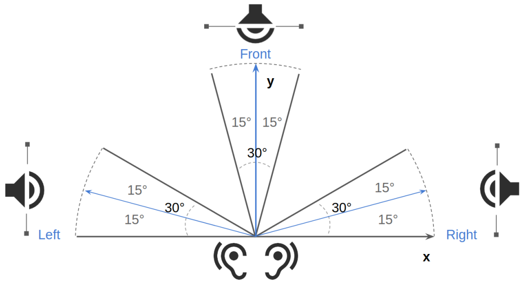

Figure 8. An illustration of the setup for the 3D Audio Experiment. “Speaker” and “ear” icons credit to Google.

5.2. 3D Audio Experiment

In this experiment, we wish to find how accurately a human can perceive the location of a computer-generated audio signal through stereo headphones. We set up this experiment so that a computer program first randomly selects one of three horizontal directions: front, left, and right. When a direction is selected, an exact angle is further generated by drawing from a uniform distribution over $\pm 15^\circ$ from the mean angle of that direction (15° for "right", 90° for "front", and 165° for "left"). A graphical depiction of this process is shown in fig. 8. Then, the program creates a virtual 3D audio source with an azimuth of the angle previously chosen, an elevation of 90° (parallel to the ground), and a radial distance of $1m$. The virtual audio source plays a 440Hz sine wave for 1.0 second, physically transmitted through a pair of headphones worn over the ears of the test subject. After the signal plays, the test subject enters the perceived location of the audio source on the computer program, after which the next virtual audio source is created and the process repeats. We gather 50 data points, each consisting of the source direction generated by the computer and the source direction perceived by the human. The confusion matrix is visualized below (see fig. 9). We include a video segment in the supplementary materials demonstrating the test program and the computer-generated 3D audio. If desired, the reader can experience a demo of the 3D audio experiment by playing the video segment with stereo headphones on.

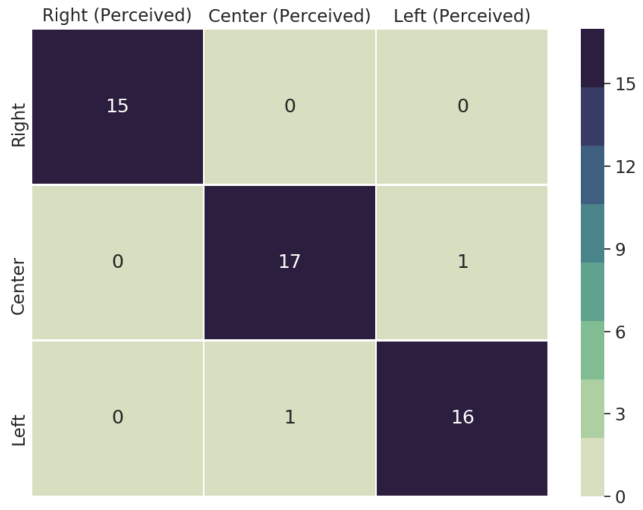

Figure 9. Confusion matrix for the 3D Audio experiment. Rows are indexed by computer-generated directions while columns are indexed by human-perceived directions.

As the data in fig. 9 shows, a human is fully capable of accurately discerning the direction of a computer-generated 3D audio signal when they are separated by different azimuth angles. In only 2 out of 50 trials have the human subject incorrectly determined the direction of the generated signal, which translates to an accuracy of 96%. This provides the insight that the 3D Audio module is a reliable way to encode spatial information into a sound to allow a listener to determine its virtual location. Spatial sounds are also highly intuitive, as almost all people with regular hearing use their two ears to locate sound sources in real life for various purposes and 3D Audio is designed to simulate this experience.

5.3. Audi-Exchange Experiment

The Audi-Exchange Experiment integrates both Hand Tracking and 3D Audio to serve as the Audi-Exchange assistive device prototype. This experiment (fig. 10) uses the "Regular Lighting / Empty Hand" images that were previously used in sec. 5.1, but we further split them into three classes, front, left, and right, based on the location of the hand appearing in the image. In each trial, an image is randomly drawn from the three directional classes and fed into the hand detector. After that, a virtual audio source is created at ground level and $1m$ in front of the user with a horizontal location being proportional to that of the detected hand such that the virtual horizontal location is within $[-1m, 1m]$. Then, the virtual location is normalized to have a distance of $1m$ so that all audio signals will have equal volume regardless of the location. In other words, if $x, y \in [0, 1]$ is the horizontal and vertical location of the hand on an image, then the location of the virtual sound source (x', y', z') is computed as:

$$\begin{aligned}(\hat{x}, \hat{y}, \hat{z}) &= ((x - 0.5) * 2, 1, 0)\\ (x', y', z') &= \frac{(\hat{x}, \hat{y}, \hat{z})}{||(\hat{x}, \hat{y}, \hat{z})||}\end{aligned} \tag{2}$$

where positive x' is the right, positive y' is the front, and positive z' is the top. Note that the vertical location of the hand is disregarded because we think horizontal location is the most important piece of information to facilitate hand-based interactions. We place the virtual audio source at (x', y', z') as computed in eq. 2, which plays a 440Hz sine wave sound for 1.0 second through headphones. Then, the test subject enters the perceived direction of the hand based solely on sound and proceeds to the next trial when ready. If the hand tracker fails to detect the hand, the program automatically records a "not detected" response. For each trial, we record the ground-truth direction of the hand and the perceived direction of the hand. The data is showed in the confusion matrix in fig. 11. The Audi-Exchange Experiment is more challenging than the independent 3D Audio and Hand Tracking experiments for two reasons. (1) Hand tracking is not perfectly accurate, so the computed image location of the tracked hand can be off. (2) When mapping the image location into the 3D location for the virtual audio source, the virtual location spans only 90° horizontally as opposed to 180° (the case for the 3D Audio Experiment) in an attempt to reproduce the actual azimuth direction of the hand in the real world, so there are smaller perceptive

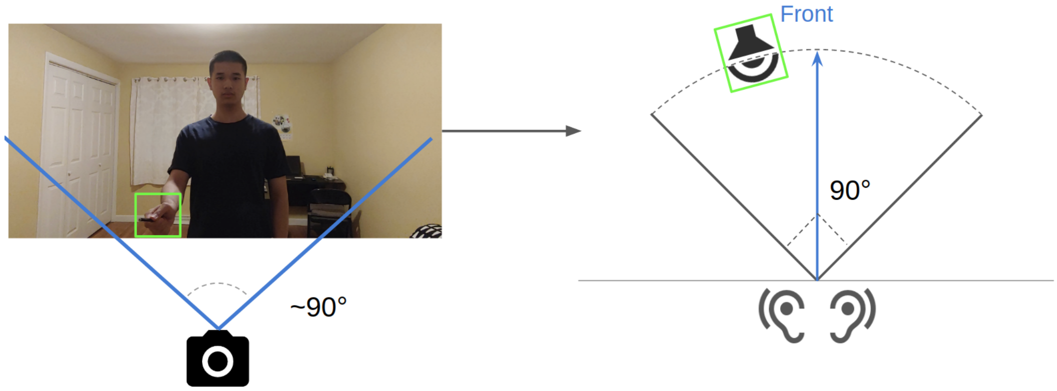

Figure 10. Illustration of the Audi-Exchange Experiment. The hand-tracking and 3D audio algorithms translate visual interaction cues of an extended hand into a 3D audio signal for the BVI user. Camera image shown was cropped to better fit the page. “Camera”, “speaker”, and “ear” icons credit to Google

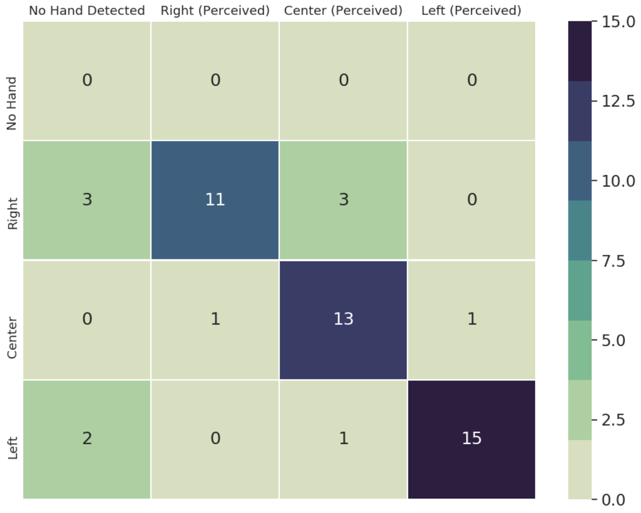

Figure 11. Confusion matrix for the Audi-Exchange Experiment. Rows are indexed by ground-truth directions while columns are indexed by human-perceived directions.

differences among sounds coming from different virutal directions. Despite these challenges, the test subject has correctly located the hand 78% of the time based on sound alone. The high tested effectiveness of the Audi-Exchange prototype strongly indicates that Audi-Exchange would be feasible and effective as an assistive platform when used by BVI individuals.

6. Porting to Mobile Devices

Although we tested the Audi-Exchange prototype on a desktop computer, it can be updated and ported to mobile computing platforms (e.g. smartphones) with relative ease. The current hand tracking algorithm is already based on a neural network architecture designed for mobile usage [15] can be run in efficiently on smartphones. While the OpenAL [3] library used in this paper is mostly used for desktop applications, newer and more powerful spatial open-source sound APIs exist that target mobile applications specifically. Resonance Audio is an example [18]. Hence, we expect porting the Audi-Exchange prototype to smartphones to be a simple process, from where we can continue to develop the Audi-Exchange assistive technology.

7. Conclusion

We have developed a working version of Audi-Exchange, a mobile assistive technology that guides the BVI when handing over and receiving objects from another person with 3D sound cues. We have implemented Audi-Exchange with a mobile-optimized neural network (i.e. MobileNet + SSD [15, 17]) and OpenAL [3]. We tested the system with images of a person using their hand to offer and receive objects captured in front of a smartphone camera and demonstrated that under proper lighting, the hand can be reliably tracked and the user can accurately determine its location by listening to a corresponding computer-generated 3D audio signal through stereo headphones. This gives us confidence that blind-or-visually-impaired individuals can use Audi-Exchange to better complete object exchanges with their hand. As the next step, we plan to construct more in-depth experiments as well as update Audi-Exchange to further improve and validate Audi-Exchange.

Acknowledgement

The authors gratefully acknowledge the financial support from the NYUAD Institute (Research Enhancement Fund - RE132).

References

[1] Google glass. 2

[2] Microsoft hololens. 2

[3] Openal: Cross platform 3d audio. 4, 8

[4] Seeing ai. 2

[5] Acesight - low vision eletronic glasses, July 2020. 2

[6] esight 3 — low vision device for the visually impaired, July 2020. 2

[7] Irisvision — wearable low vision glasses for visually impaired, May 2020. 2, 3

[8] Orcam myeye 2.0, July 2020. 2, 3

[9] Pioneering accessible navigation – indoors and outdoors, May 2020. 2

[10] Isaac Attia and Daniel Asamoah Brempong. The white cane. its effectiveness, challenges and suggestions for effective use: The case of akropong school for the blind. *Journal of Education, Society and Behavioural Science*, pages 47–55, 05 2020. 2

[11] Sachin Bharambe, Rohan Thakker, Harsharanga Patil, and K. Bhurchandi. Substitute eyes for blind with navigator using android. pages 38–43, 04 2013. 2, 3

[12] Wafa Elmannai and Khaled Elleithy. Sensor-based assistive devices for visually-impaired people: Current status, challenges, and future directions. *Sensors*, 17(3):565, 2017. 2

[13] Leire Francés, P. Morer, Mabel Rodriguez, and Aitor Cazón-Martín. Wearable design requirements identification and evaluation. *Sensors*, 20:2599, 05 2020. 2

[14] RE Hoover. The cane as a travel aid. *Blindness*, pages 353–365, 1950. 2

[15] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 3, 8

[16] Vladimir Kulyukin, Chaitanya Gharpure, John Nicholson, and Sachin Pavithran. Rfid in robot-assisted indoor navigation for the visually impaired. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 2, pages 1979–1984. IEEE, 2004. 2

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016. 3, 8

[18] Mathew Block Martin Dufour, Aaron McLeran et al. Resonance-audio, 2018. 4, 8

[19] João Fradinho Oliveira. The path force feedback belt. In *2013 8th International Conference on Information Technology in Asia (CITA)*, pages 1–6. IEEE, 2013. 2, 3

[20] BR Prudhvi and Rishab Bagani. Silicon eyes: Gps-gsm based navigation assistant for visually impaired using capacitive touch braille keypad and smart sms facility. In *2013 World Congress on Computer and Information Technology (WCCIT)*, pages 1–3. IEEE, 2013. 2

[21] Uta R Roentgen, Gert Jan Gelderblom, Mathijs Soede, and Luc P De Witte. Inventory of electronic mobility aids for persons with visual impairments: A literature review. *Journal of Visual Impairment & Blindness*, 102(11):702–724, 2008. 2

[22] Roy Shilkrot, Jochen Huber, Connie Liu, Pattie Maes, and Suranga Chandima Nanayakkara. Fingerreader: A wearable device to support text reading on the go. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, page 2359–2364, New York, NY, USA, 2014. Association for Computing Machinery. 2, 3

[23] Dibia Victor. Handtrack: A library for prototyping real-time hand trackinginterfaces using convolutional neural networks. *GitHub repository*, 2017. 3, 4, 6

[24] Mohd Helmy Abd Wahab, Amirul A Talib, Herdawatie A Kadir, Ayob Johari, Ahmad Noraziah, Roslina M Sidek, and Ariffin A Mutalib. Smart cane: Assistive cane for visually-impaired people. *arXiv preprint arXiv:1110.5156*, 2011. 2

[25] Wikipedia contributors. Head-related transfer function — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Head-related_transfer_function&oldid=940911588, 2020. [Online; accessed 18-July-2020]. 4