AWARD NUMBER: W81XWH-21-1-0615

TITLE: Delivering Sensory and Semantic Visual Information via Auditory Feedback on Mobile Technology

PRINCIPAL INVESTIGATOR: Kevin C. Chan, Ph.D.

CONTRACTING ORGANIZATION: New York University Grossman School of Medicine

REPORT DATE: OCTOBER 2023

TYPE OF REPORT: Annual Report

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED**
OCTOBER 2023	Annual Report	*1SEPT2022 - 31AUG2023*

4. TITLE AND SUBTITLE

Delivering Sensory and Semantic Visual Information via Auditory Feedback on Mobile Technology

5a. CONTRACT NUMBER
W81XWH-21-1-0615

5b. GRANT NUMBER
W81XWH-21-1-0615

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kevin C. Chan, PhD and Giles Hamilton-Fletcher, PhD

Emails: Kevin.Chan2@nyulangone.org; Giles.Hamilton-Fletcher@nyulangone.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

New York University School of Medicine
Department of Ophthalmology
222 East 41st Street, 3rd Floor
New York, NY 10017-0000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command

Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project seeks to create and assess new visual-assistive smartphone Apps for fully blind end users to better interact with the visual environment. These Apps convey sensory information gathered by cameras/sensors (e.g. color, distance, heat) and semantic information from artificial intelligence (e.g. object identity, shape, size, location in the image). This information is conveyed through spoken verbal feedback (e.g. "chair, bottom left; TV middle right") and/or 'musical' audio (e.g. musical 'meows' that play from a cat's location). Our research purpose is to produce new Apps that increase visual information accessibility, enhance daily functionality, and facilitate new interactions of interest to blind end users. In terms of scope, this 2-year project focuses on the initial development of novel technologies in the first year, with at-home beta-testing by fully blind subjects and further technology refinement in the second year. In year 1, we combined the modern iPhone's 3D sensors (e.g. LiDAR range-finding) and DeepLabV3 object segmentation, to provide the required information, all running in real-time, locally on iPhone. By providing objects and their distances, we provide a stable and intuitive understanding of the environment that remains consistent across variable lighting or minor changes in object features. Overall, by combining object identity, distance, and their visual features effectively, this goes beyond prior technologies in providing both sensory and semantic-level information to the user, either separately, or in a novel combined 'hybrid' format. In year 2, we provided blind beta testers with our Apps and supporting technology to try at home, followed by a series of interviews and questionnaires to assess their experiences of the Apps, suitability for daily tasks, and their desires for future technologies. Based on our findings, we are currently revising our Apps to meet their requests. This includes expanding the range of objects recognized by the system, further optimizing how LiDAR and DeepLabV3 data is integrated, and improvements to how the final musical/spoken soundscape communicates this information. Our outcomes are expected to aid user comprehension, and improve the intuitiveness, usability, and functionality for day-to-day tasks.

15. SUBJECT TERMS
Visual Assistive Technology; Sound-Vision; Computer vision; Visual Rehabilitation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRDC
U	U	U	UU	**23**	**19b. TELEPHONE NUMBER** *(include area code)*

TABLE OF CONTENTS

1. INTRODUCTION:

'Sensory substitution devices' (SSDs) are assistive technology apps that convert visual images into patterns of sound, enabling blind listeners to hear the distribution of color/heat/depth in an image. Despite their rehabilitative potential, these *'sensory-level'* devices are rarely adopted by the blind community due to their initial impracticality and long learning phases. Modern computer-vision object recognition techniques can address these issues by providing *'semantic-level'* content (e.g. object name), as well as interacting with SSDs to provide a *'hybrid'* experience, such as hearing the shape, size, location, and identity of objects. This project will create and provide Apps for all 3 levels of information to subjects with blindness for safe at-home user testing and feedback. User feedback will be used to refine the App, and be ready for public release at the end of the study.

2. KEYWORDS:

Visual Assistive Technology; Sound-Vision; Computer vision; Visual Rehabilitation

3. ACCOMPLISHMENTS:

What were the major goals of the project?

The major goals of the project are to develop an assistive technology smartphone application for persons with blindness (Specific Aim 1) that conveys:
(1) 'Sensory information' (e.g. distances, colors, shapes via musical tones – Major Task 1)
(2) 'Semantic information' (object identities via spoken feedback, e.g. "person" – Major Task 2)
(3) A sensory and semantic 'hybrid mode' (e.g. communicating objects and their visual properties using musical tones and/or speech – Major Task 3).

This application is then beta-tested by blind end users (Specific Aim 2). This requires HRPO and IRB approval (Major Task 4), as well as 3 batches of 10 blind subjects being recruited, providing feedback, and being interviewed (Major Tasks 5-7), with this information being written and published (Major Task 8).

Timeline and milestones are reported below this box.

Timeline / milestones

Throughout this report we will discuss a variety of distinct modes, terms, and features across several Apps. Each of these falls under either the 'sensory', 'semantic', or 'hybrid' umbrella terms that make up the core of this research. Below we provide a key that will help clarify how each mode, term, or feature relates to our 3 core terms:

Key Terms:

- **Sensory:** Sensory-mode, brightness, color, distance-to-music, distance-to-sound, distance-to-vibration. The SoundSight app and light detector app.
- **Semantic:** Semantic-mode, spoken, verbal, verbal-mode. AI-Sight's "objects" and "live-objects" modes.
- **Hybrid:** Hybrid-mode, objects-to-music, objects-to-sound, music-mode. AI-Sight's "play" and "live-play" modes.

Our SoundSight App provides only sensory information as it provides either distances or colors as musical feedback. The AI-Sight App features modes that correspond to semantic information ("objects", "live-objects") and hybrid information ("play", "live-play"). Discussions around AI-Sight's 'live-modes' refer to the modes which constantly provide spoken or musical feedback to the user during use ("live-objects" and "live-play") while their non-live equivalents ("play", "objects") refer to modes that the user has to manually select using a button for them to provide their feedback.

Specific Aim 1: Develop prototype smartphone Apps to convey sensory and semantic information.

Major Task 1 (Months 1-3): Develop and implement methods to convey basic sensory features (color, distance) from smartphone cameras / sensors via auditory feedback. While 90% of originally specified features are complete, all essential features are completed so it is suitable (and has been used) for beta-testing.

Given the starting month of September 2021, the timeline lists Major Task 1 (sensory-mode) as having a target date of January 2022. For subtask 1, accessing depth and color information via intuitive audio feedback is done via our original "SoundSight" App platform for the user testing – here users have access to different forms of vision-into-audio conversion. This App has been refined with a variety of curated modes suitable for testing in this study. During this development, we experienced repeated incompatibility issues with the plug-in Flir One thermal cam and so thermal image sonification is not currently supported in SoundSight. On the other hand, in this 2023 reporting period, we have overcome the previous technical challenges, and successfully integrated LiDAR (Light Detection and Ranging) distance information into our main App, the "AI-Sight". Work is ongoing to ensure the best possible combination of distance information with object segmentation AI in the App, and we detail this process in our accomplishments for this reporting period. We will continue to refine its integration based on user feedback from the 1st batch of research subjects in this reporting period, while the updated version of the AI-Sight App will be introduced to the 2nd batch of research subjects (**Subtask 1, 95%**).

For subtask 2, we previously explored on plane detection (segmenting floors, walls etc) using Apple's ARKit platform. We have the ARKit codebase working within the SoundSight App, which provides stable distance measurements and estimations using LiDAR. Further development on adding plane detection functionality for this subtask 2 was on hiatus due to our use of DeepLabV3 AI which has already eliminated walls/floors in our semantic and hybrid modes. As stated in prior reports, this reduces the importance of this subtask, leaving it as a low-priority future development feature. That said, our continuing work in training DeepLabV3 to recognize new objects of interest also includes segmenting floors and walls, and hence this AI may instead provide us the option to 'turn on or off' whether walls / floors should be reported back to the user, which may be a superior approach than our originally envisioned ARKit approach. Finally, the addition of LiDAR distance estimation to the AI-Sight allows us to provide audio alert on objects close to users (**Subtask 2, 95%**).

In relation to exploring various sensory-level options, in this 2023 reporting period, we evaluated five different approaches to distance estimation on smartphones (ARKit [front-facing, back-facing], LiDAR, Infrared grid distortion, and machine learning approaches). Our evaluation on the accuracy and usability of these approaches was written up as a research paper which was submitted to the IEEE Open Journal of Engineering in Medicine and Biology (IEEE OJEMB). This was peer reviewed and had minor revisions requested. The revised manuscript was submitted back to the journal and is currently under review. This contributes to **Milestone 3** of this project by dissemination of initial design/results via publications.

Major Task 2 (Months 4-8): Develop and implement computer-vision technologies in our AI-Sight smartphone App that recognizes semantic-level content (i.e., objects in the environment) and provides this information as verbal feedback to the user. 95% of features complete, ready (and used) for beta-testing.

Major task 2 (semantic-only mode – within the "AI-Sight" App) has a timeline of Feb-May 2022. The AI-Sight's semantic-mode can track the location and identity of multiple objects simultaneously within an image (**Subtasks 1 & 5, 100%**) and convey the horizontal and vertical positions of these objects via descriptors and/or vocal-pitch (**Subtask 3, 80%** - not 100% due to lack of spatialized audio for verbal feedback). The position and identity of multiple AI-recognized objects can update in real-time at ~8 frames per second (**Subtask 4, 100%**). Accessing semantic-modes can be done via a 'double-tap' of the screen, while switching to representations of other levels is done via App-switching to the SoundSight (for sensory-modes), or via a long-press on the screen (for hybrid-mode). Since Dec 2022, we also introduced an on-screen 'semantic-mode' button (called "objects") and a 'hybrid-mode' button (called "play"), the purpose of which is to work with Apple's VoiceOver screen reading software to increase accessibility (**Subtask 7, 90%**). In terms of integrating additional sensory information, in April 2023 we managed to add LiDAR distance information to the AI-Sight App, and as of November 2023 we have distance values extracted for all objects present with pixel-level precision that is usable for both semantic and hybrid-modes (**Subtask 2, 100%**). This distance integration was prioritized for development over thermal information, due to prior research indicating that distance is a higher priority for users who are blind, and the general inaccessibility of the thermal cam supporting hardware for blind users (**Subtask 6, 0%**). It should be noted that user testing can progress regardless of the presence or absence of thermal information. **Relevant to subtasks 1 & 5**, we published and presented a paper at IEEE EMBC 2023 in July 2023 on training a YOLOv5x AI object recognition algorithm to recognize 35 objects of interest to the blind and visually impaired community, which contributes to **Milestone 3**. This work provides major future directions for our Apps to recognize specific key objects for visually impaired users. Building on this, since March 2023, we have also performed an initial training of our DeepLabV3 AI to recognize additional object categories related to instrumental activities of daily living with a view to implementing them into the AI-Sight App in the future. Since September 2023, we have also further trained a new object detection AI (YOLOv8) to recognize new object categories for persons who are blind (e.g. doors, door handles) on top of its 80 object baseline. We plan to submit work on training DeepLabV3 and YOLOv8 to IEEE EMBC 2024 and OJEMB, which contributes to **Milestone 3**.

Major Task 3 (Months 9-12): Develop a 'hybrid mode' that provides a unique combination of both sensory and semantic-level information within a smartphone App. 90% of features complete, ready (and used) for beta-testing.

Major task 3 (hybrid-mode – within the "AI-Sight" App) has a timeline of June-Sept 2022. This mode segments multiple objects from images and provides their sensory features (location, size, shape, distribution in the image) as well as semantic information such as object identity and location (**Subtask 1, 100%**). This information is presented in multiple ways to the user: a 'long-press' of the screen or pressing the "play" button activates 'hybrid-mode' (see figure 1) which uses musical feedback to convey the distribution of objects, their size, shape, and identity to the user, while a 'double-tap' on the screen, or pressing the "objects" button conveys the 'semantic-mode' for which each object's position and identity are provided via spoken feedback. We have also introduced two 'live-modes' for users (see figure 2), where if enabled, rapidly and constantly provide a streamlined version of semantic and/or hybrid feedback to users informing them of the object locations and identities, without the need for direct user interaction (**Subtask 2, 100%**). Each style of information or feedback can be enabled at any time by the user, or 'cued-up' to present the same image in two formats one after another. This can be done for both the more simplistic 'live-modes' and more in-depth user-selected modes, allowing both verbal and musical feedback simultaneously in order for multiple levels of image representation (semantic, hybrid) to be communicated to the user at a given moment (**Subtask 3, 100%**). We are currently working to also include purely sensory-level information (e.g. distance feedback) so that the user can go between all 3 levels of information (sensory, semantic, or hybrid) all within the AI-Sight App.

Since April 2023, we continued to develop and refine the 'live-view' modes (both semantic/verbal, and hybrid/musical) into an easy-to-use mode for end users. Users are able to toggle these modes during App use, with the semantic-level 'live-objects' reading out central object names, and the hybrid-level 'live-play' which constantly scans the whole image for objects and communicates via musical feedback. These live-view modes provide constant feedback about the contents of the image in a condensed format to provide general contextual awareness to the end user. This helps users understand *when* there is content to investigate further using the button-selected semantic or hybrid-mode feedback. We have fine-tuned how and when this information is

verbalized through combining (1) the use of selecting the center-most object, (2) using more conservative size-thresholding to avoid small misrecognitions in the image, and (3) intuitive moments to refresh the object list. These live-play and live-object versions were created for the first batch of user testing. In response to end user feedback, the updated version created for the second batch of user testing will also feature several new refinements. This includes: new object prioritization weighting (size, centrality, distance); smarter timing for when updates are provided to the user (e.g. updating the user when either the central object or its distance has changed); additional thresholding procedures (e.g. object must be present for 3 consecutive frames to be verbalized); adding distance values to feedback to this information (e.g. "table at 3ft"); and new forms of communicating distance musically (e.g. faster beep repetition rates – similar to a parking sensor). These updates are detailed further in the 'what was accomplished section' below.

Overall, for specific aim 1, technical development has produced all essential features required for beta-testing with the initial versions of these systems already tested by the first batch of beta-testers. We have ensured that these systems are accessible from App startup (the live modes) while all options are accessible using Apple's VoiceOver screen reading software (the button modes) so that all features are easily available for users to try out and provide feedback on. The technical development and integration of LiDAR distances with DeepLabV3 object segmentation have meshed well, ascribing distance information appropriately to each recognized object. We have also pushed ahead on publications in this space, with one on AI-training for IEEE EMBC, one on distance estimation techniques for IEEE OJEMB, and two to three more papers in preparation relating to AI-training (**Milestone 3**). As described above, the underlying systems in the AI-Sight are still being refined but are more advanced than was provided for the first batch of testers. Once the system is fully updated it will be suitable to write up a technical description of the system and its functionality for publication (**Milestone 10**).

Specific Aim 2: Conduct observational research on how the App performs during beta-testing by blind end users.

Major Task 4 (preparation & recruitment), has a timeline of April to June 2022 for document approval for human testing. The original documents, written at the start of the project, were originally approved by Apr 2022 by both the USAMRDC Human Research Protection Office (**Subtask 1, 100%**) and NYU Langone Health Institutional Review Board (**Subtask 2, 100%**). Since the submission of the updated materials within this reporting period, confirmation of approval from NYU Langone Health Institutional Review Board was subsequently given on March 2nd, 2023, and approval from the USAMRDC Human Research Protection Office was provided on March 21st, 2023. In terms of enrollment, we have recruited our first 9 subjects for our first batch of user testing (**Subtask 3**, **40%**). Our list of potential subjects has also expanded due to word-of-mouth within the visually impaired community originating from our beta-testing subjects. We will be working to supplement this with refreshed subject recruitment lists during the no-cost extension period.

Major Task 5 (beta-testing by blind end users), 9 blind subjects were recruited and consented into the study, where iPhones and Apps were provided (**Subtask 1, 100%**), each subject was provided with sensory-only, semantic-only, and hybrid-mode experiences with the Apps and subsequently underwent a remote midway interview on their experiences with the Apps (**Subtasks 2, 3, and 4, 100% each**). Subjects have also conducted their final interview and series of questionnaires as well as the return of provisions (**Subtask 5, 100%**). Revisions to the AI-Sight technology based on their interview and questionnaire feedback have begun with several major changes to the underlying information processing in the App (e.g. LiDAR integration with DeepLabV3 at the pixel level) done in order to support their preferred methods of feedback such as verbalization and regular updating of distances to objects; 'parking sensor'-like audio feedback (**Subtask 6, 70%**). This marks completion of beta-testing for batch 1 of subjects (**Milestone #7**). After Major task 5 is complete, we will start on **Major Tasks 6 and 7** with the revised AI-Sight for batches 2 and 3 of beta-testing subjects, with subsequent revisions based on their feedback.

Major Task 8 (Final technological refinements and dissemination of results), with our first analysis of beta-testing feedback and initial revisions to the Apps, we have started progressing on final technological refinement (**Subtask 1, 30%**) and data analysis and summaries (**Subtask 2, 30%**).

What was accomplished under these goals?

Please read below our goal reporting for this period.

Technical Development (specific aim 1)
During this reporting period, we continued the technical development of the SoundSight and AI-Sight Apps (specific aim 1) so that the sensory, semantic, and hybrid-modes were in a suitable place for initial beta-testing. In addition, we continued the revisions of the Apps in response to feedback from the first batch of beta-testing.

SoundSight App progression (Sensory-mode)
To allow beta-testers to explore using sensory-level feedback, we updated the SoundSight App's approach to distance estimation and implemented curated sonifications. For distance estimation, we implemented the ARKit framework as the method for providing our LiDAR-derived depth map rather than AVDepthData, which had previously constrained us to an overly narrow field-of-view. We produced our public-facing user interface for the App that we ensured would work with Apple's VoiceOver accessibility interface. Our user interface allows users to select between nine different distance/color-to-sound modes, each of which contains a description of the mode they are about to hear. This progresses **Milestone 10**. While SoundSight will be used for all rounds of beta-testing, further sensory-mode development will concentrate on implementation within the AI-Sight App.

AI-Sight App progression (semantic-mode, hybrid-mode) for beta-testing group 1
In preparation for the first beta-testing group, we made several refinements to the available modes that convey semantic or hybrid content, providing users with gesture/button-selected modes that provide more detail (figure 1), and live 'always on' modes that provide more convenient and accessible summaries of the environment (figure 2). The buttons are called "Objects" for semantic-mode, and "Play" for hybrid-mode, with these names chosen to be as short as possible, so VoiceOver's speech for the button names does not overlap with the spoken feedback from the App itself. Descriptions on how each mode functions are featured within the figure descriptions.

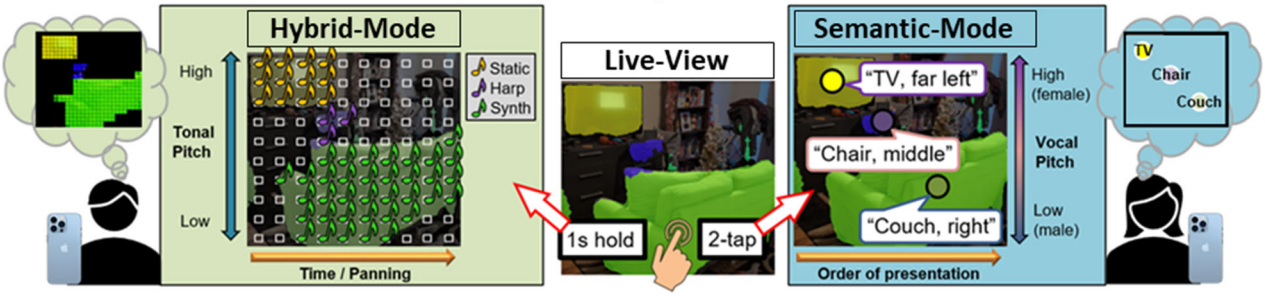

Figure 1. AI-Sight App's button-selected modes. The App *segments* recognized objects from live visual images on iPhone (middle image; 'Live-View'), these objects are color-coded (yellow = TV, couch = green, chair = purple). If the user holds a press on the screen for 1 second or presses the "play" button (left image; 'hybrid-mode'), musical notes are played for all pixel locations (white boxes) that contain recognized objects. Musical notes are played column-by-column over time from left-to-right, while spatially panning from left-to-right. Higher pixels play higher pitched notes, while the object in the pixel determines the notes' sound quality (e.g. couch = synth, car = horn). Higher pixels are also played slightly before lower ones in each column, creating a high-to-low pitch strumming effect. The playing of each column is preceded by a subtle 'drum' effect, which allows the user to track the 'scan-through' even if no objects are detected and it would otherwise be silent. Over time, these notes form the structure and content of the visual image in the mind of the blind listeners. If the user double-taps the screen or presses the "objects" button (right image; 'semantic-mode'), the recognized objects' names are read-out from leftmost-to-rightmost, with the horizontal position described, and the vocal pitch of the speaker indicating object height.

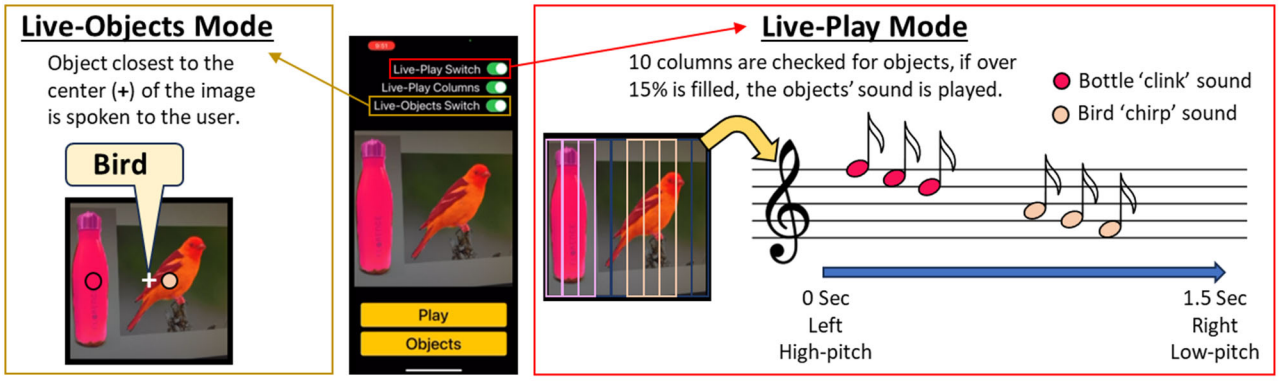

Figure 2. AI-Sight App's live-modes. Currently, users have the option of toggling on two live modes that constantly provide feedback to the user while the App is running. In the left image, live-objects mode is enabled, which provides constant semantic-level feedback. Here the object that is closest to the center of the image is verbally spoken to the user. If this status changes, such as a different object becomes the centermost object, this is read out to the user. Alternatively, if no object reaches the size threshold (>10% of the image), then the next object to reach this size-threshold is read out to the user. In the right image, live-play is enabled, which provides constant hybrid-level feedback. Here the image is subdivided into 10 columns of equal size, and each column is checked for objects. If an object takes up over 15% of the column, this queues up the appropriate sound (sound quality, laterality, pitch, timing) for playback during the left-to-right scan over 1.5 seconds. For example, in the displayed image, the bottle takes up 3 columns, and so schedules 3 bottle 'clink' notes for playback, the gap between the bottle and bird is heard as a period of silence, and then the bird (which takes up 3 columns) is heard as 3 notes of bird chirps, followed by silence for the remaining columns with no objects in them. For this example, the overall soundscape is heard as 3 bottle 'clinks', brief silence, 3 bird 'chirps', brief silence.

The constantly running live modes and button-selected modes have distinct purposes for the beta-testers. Here the live modes provide context for the user, instantly alerting them to basic information about the environment, such as the presence and location of objects. This provides general environmental awareness and supplements the button-selected modes that provide the option for more detail, such as the naming of all objects and descriptions of their locations in the image, or an extensive musical playthrough that 'sketches' the distribution of objects (and their size, shape) in the image, allowing the user to mentally reconstruct the image in their mind. Of note, since we cannot control whether our beta-tester are using headphones or listening to the smartphone speakers, we communicate important information using several auditory features (e.g. left-right position via timing, spatialization, pitch), this allows beta-testers access to the same information irrespective of their method of listening. In addition, to provide robust, stable feedback to the end user, we implemented size-thresholding to the "objects" button, live-objects, and live-play modes. This prevents communicating small misrecognitions by the DeepLabV3 AI system; however, we have also worked to develop this area further to avoid larger but 'brief' misrecognitions (see below). The AI-Sight version provided to the first group of beta-testers did not have distances integrated into the system. This was developed after the first round of beta-testers were recruited, as such, their feedback mainly reflects the semantic and hybrid-modes for the AI-Sight App, whereas distance feedback was provided from our SoundSight App used during the beta-testing.

User Testing (specific aim 2)

For our beta-test, we recruited nine blind subjects (3 female, mean age = 56.9 years, S.D. = 13.2). After completing the consenting process, subjects were provided with an iPhone to support the use of the Apps, and we either confirmed or taught users to operate the iPhone and navigate to the Apps using blind accessibility modes (Apple's VoiceOver). For each App, subjects were taught all the options inside them and allowed to freely explore the research room. This typically involved subjects identifying the location and identity of objects within the room (e.g. person, chair, bottle, table, TV) using the unique feedback of each App. Subject questions were answered regarding each App. Subjects were informed of the general nature of the interview questions they would be asked, and reminded to only use the Apps within safe situations (i.e. To only use in safe circumstance, with examples provided; where the successful or unsuccessful use of the App has no effect on their safety; To not have the Apps distract or disrupt from other safety behaviors or their orientation & mobility training). Subjects were interviewed via phone call after 2-3 weeks of device usage, and during their final visit after 1-3 months of device usage they were interviewed again alongside being asked to rate each mode (sensory, semantic, hybrid) on a variety of questionnaires. One subject dropped out of the study around this time owing to changes in life circumstances that meant he felt they could not devote appropriate time to the study and returned the iPhone that was provided to him.

The midpoint and final interviews provided general use themes and App feedback. These are briefly summarized below, with specific feedback from our nine participants numbered from p1 to p9.

Where was the App used? Subjects tried out a variety of locations, including exploring their room/house/apartment, building patios, walking in the walk, visiting a relative's house/backyard, observing while at an airport, restaurant and a supermarket.

What did users think of sensory-level information (brightness, color, distance)?

- *Light information* – This was well liked by some (p3, p7) but regarded as useless by others (p4, p5). This was primarily used to check whether lightbulbs were on or off, and the simplicity of feedback was appreciated, while for some user's lightbulb checking was not a concern and so perceived this information as pointless.
- *Color information* – This was enjoyed by some subjects as practical for clothes, boxes, and color-coded objects (p7), but considered too unreliable (p3) or confusing (p8) by others.
- *Distance information* – Subjects unanimously had positive feedback for this in the Apps, noting it was reliable, interesting, worked in poor lighting conditions, was a good backup for when object recognition could not help or was too slow at updating, essential for exploring/navigating, "distance is the best function the App has."
    - *Comparing feedback approaches:* relative to spoken feedback (e.g. "table 8ft away"), changes in auditory loudness, tempo, or frequency, helped users contextualize distance descriptions and how the gap between the user and object is closing as they walk towards it. Some users described verbal descriptions of distance as confusing in isolation - e.g. words to the effect of "how far is 32ft? I don't know what this means, how many steps is that?", while constantly updating auditory intensity is easier to track distance changes as they get closer to the object. "I liked the use of the rapid tone that got faster the closer I was to the door (p1)".
    - *Potential psychological* effects. Sensory-level feedback that regularly updates may help provide a feeling of connection between the user and object, in a style that may help egocentric navigation that is commonly used in the blind community.

What did users think of the feedback for sensory, semantic, and hybrid modes?

- SoundSight (**Sensory-mode**, distance-into-musical audio):
    - This musical feedback was praised for its responsivity in communicating distance, with simpler modes that only use variations of a single style of instrument receiving the highest praise. Many subjects enjoyed that closer objects were louder in their explorations. However, more complex audio modes that use multiple sound qualities were considered both harder to use, and the overall sound quality feeling more 'muffled'. The biggest concern came with the perceived learning curve, attention required, which could make the experience more stressful. This led one subject to want to quit this mode (p6), but others were happy to persist "once you understand the sound and adapt, it's very helpful" (p5). The left-right sweep of information (which is a very common approach in designing these devices) was also considered confusing by some. The attention required for the more complex soundscapes could take away from users perceiving their surroundings (hence simplicity is key), and as expected it was generally considered hard to recognize objects without AI assistance.
- AI-Sight (**Semantic-mode**, objects-into-speech).
    - Information on types of objects was considered good for unknown spaces, and it was considered the clearest feedback in our Apps. However, it was limited in only being able to recognize 20 object categories in the current version, making usage rather repetitive and thus hoping to enable recognizing more object categories. Also while it may tell you the presence of objects, it does not tell you the objects distance (in this version), your angle towards the object, or any objects in the way which is important to help navigation. Some misrecognitions were surprising to users, and hence some object thresholds may need to be further calibrated. The spoken feedback can be slow in comparison to the sensory/hybrid-level feedback and hence requires a failsafe (such as distance-to-audio) in case objects are not recognized by the AI system in time. The ideal version of semantic feedback was widely considered to be object + distance (either in feet, or steps to object) and potentially direction as well e.g. "table at 3ft, on your left" as it strikes the right balance between brevity and key information. The left-right verbal descriptions were liked for observing scenes.
- AI-Sight (**Hybrid-mode**, objects-into-musical audio)

- This was considered good for looking for nearby objects (such as those on the floor) or general awareness of the environment, especially since the App had a left-right distinction in the feedback that was absent from most others. However, subjects frequently lamented the lack of distance information in the AI-Sight's feedback. This prevented the App from being a 'navigation' device to instead being an 'awareness' device, as the information could not be easily acted upon outside of reaching distances. "AI-Sight… didn't give out distances, lacking 'door & 3ft' so felt lost, couldn't *navigate* well with it, so it's more for sitting down." (p1). There was a preference towards simplicity so that the soundscape does not get cluttered or confusing, with as few overlapping sounds as possible.

Below we discuss some take-away points brought up by the feedback so far:

Preferred Auditory Approaches. Currently the most liked auditory approach by subjects was that of increasing the tempo of 'beeps' to indicate closer objects. This is likely due to its simplicity and intuitiveness (mimicking a car parking sensor). Relevant to these findings, a recent paper on communicating distances via different forms of auditory feedback (loudness, frequency, tempo etc) found that *tempo* increases, which they refer to as 'beep repetition rate' was both the most intuitive to users and the most accurate in terms of gauging distances [1]. This may be a development path forward to ensure that users have a good understanding of object distances, however it remains to be seen how this method fares when communicating multiple object distances at the same time, and how this interacts with verbal feedback in a responsive manner.

Variations in Acceptance of Audio Feedback. Subjects were diverse in whether musical forms of audio feedback were seen as pleasant or frustrating. For some subjects, *feeling* the presence of external objects or sensory features was an enjoyable experience in of itself, irrespective of whether it facilitated a task, while for others, if the behavioral endpoint was not being directly facilitated the feedback was seen as frustrating. This leads us to think of an interesting distinction in our subjects as being either 'experiential' or 'purely task-focused'. This may be influenced by many factors that warrant further investigation, including personality traits and hobbies (e.g. musical interest). We will also investigate whether the Apps can be built and tailored for specific experiential or task-relevant purposes.

Egocentric vs Allocentric Navigation. Of note, prior studies have found that blind subjects, and especially, congenitally blind subjects, tend to use egocentric (self-to-object relations) rather than allocentric (object-to-object relations) approaches to mentally represent the environment. The type of feedback generally preferred in the Apps by our subjects helps feed into egocentric understandings of the environment (i.e. what is *your* relationship to this single object) and hence may feel more intuitive to users. This is well captured by the following quote: "I want regular updates on distance to objects such as approaching beeps/vibration to show how <u>you're</u> moving towards the door rather than its location" (p1). This is also reflected in wanting to know obstacles on the path to that object, and have this information conveyed by what corrective behaviors the user should take in navigating to the object – e.g. "Door at 8ft, chair in the way at 3ft, carve left to avoid" (p1). This approach provides 'where-to-go' as opposed to 'where-things-are' information.

Objects Requested for Recognition. As we will continue to train our AI system to recognize new objects, we were interested in what objects our subjects were interested in, here are their responses: (p1) Pillar, ramp, steps, bathroom sign, doors, checkouts, register, desk, counter, faces (facial recognition). (p3) stairs, garbage can, fences, tree, lamppost, phone, lamp, printer, stairs, bench, counter (reception desk), pen, door, elevator, ramp (or changes in ground level). (p4) Doors, curbs. (p5) crosswalks, traffic lights, garbage can, window, cabinets, kitchen (cabinet door – open vs closed), shoes, cane, keys, keychain. (p6) tree, door. (p7) Wall, screen/window, pictures. (p8) bowl, tree, window, bed, book, glasses. (p9) Book, wallet, money, images, general 'object' category (even if not identified).

Sensory vs Semantic vs Hybrid. In terms of comparing these approaches, sensory information appeared to be valued the most for enjoying specific tasks, as a consistent safety fallback and to contextualize AI-feedback but was not viewed in isolation as being especially helpful to facilitate a deeper understanding of the environment. Semantic information by itself was seen as useful as a general awareness and observation tool but not facilitating navigation behaviors, while the hybrid of both sensory and semantic information appears to progress the Apps towards a practical navigation tool that provides both environmental understanding and facilitates the subject's movement through it. Here it could be said that sensory information contextualizes semantic information, rather than the other way around.

We also have feedback on several other aspects that will be saved for future publications, including how the Apps changed the subject's behavior, other technical features requested, user interface changes, additional aspects to consider (battery,

phone/hand usage, safety, connectivity), stumbling blocks to App usage, and the subject's top requested additions (e.g. text recognition, image descriptions).

Quantitative Feedback – Questionnaire results. During the final session with the beta-testers, participants were asked to rate the SoundSight's sensory-mode ("distance-tones") as well as the AI-Sight's semantic ("objects-speech") and hybrid-modes ("objects-tones") across a series of questionnaires on usability and practicality. The questionnaires included: (i) ratings out of 10 for usability, practicality, and aesthetics; (ii) ratings between 1 and 7 in NASA-TLX for different types of demand placed on the user and overall performance; (iii) Rating the practicality of these modes out of 10 for various tasks; (iv) The AttrakDiff questionnaire where users rate where each mode lands between a series of contrasting words (e.g. isolating-connective) on a 7-point Likert scale; and (v) the System Usability Scale, where users rate on a 5-point scale their agreement with statements on a variety of usability metrics.

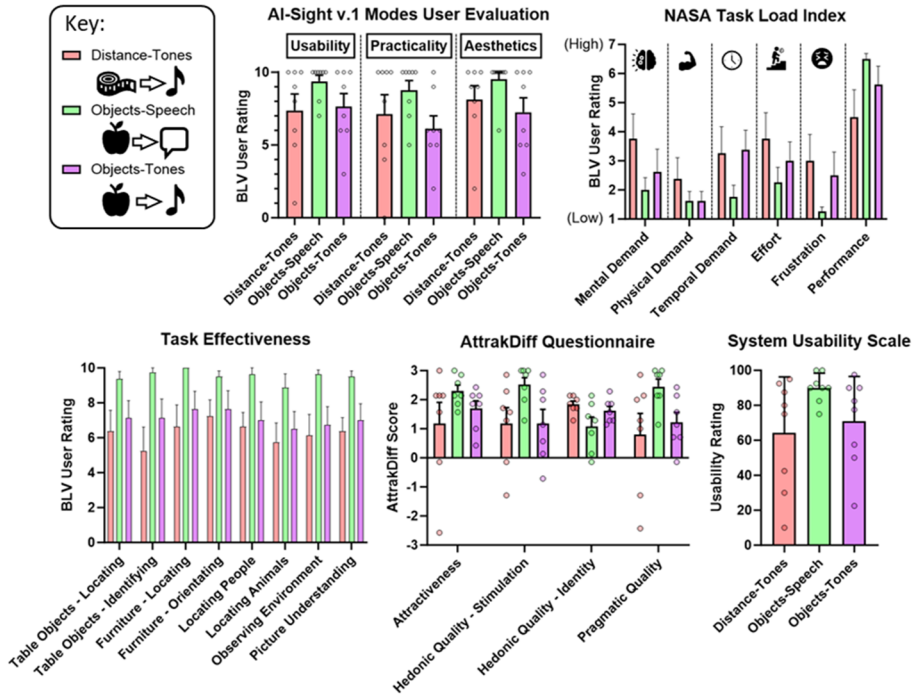

Figure 3. Quantitative outcomes by SoundSight and AI-Sight beta-testing users. Users rated specific modes from our Apps, specifically, distance-tones (red), objects-speech (green), and objects-tones (purple), which are the sensory-, semantic-, and hybrid-modes respectively. Beta-testers rated these across five questionnaires with higher scores indicating more positive ratings except for the first five scales for the NASA Task Load Index for which lower scores indicate less demand being placed on the user. Error bars show +1 standard error of the mean.

The feedback from our beta-testers on these quantitative measures revealed consistently high ratings given to object-speech (semantic-mode) across a diverse series of metrics. The conversion of distance or objects into tones received positive feedback from most users, although some preferred verbal feedback over this method. The data was analyzed using a series of ANOVAs, the exact statistics of which we will report in future publications. When considering statistically significant results, the different approaches varied in their user demand, with the object-speech method being the least taxing. Moreover, users found object-speech to be more effective for various everyday activities than the other methods, and significantly more practical than distance-tones.

Insights from our interviews and questionnaires highlighted that each of the three modes – sensory, semantic, and hybrid – offered distinct and valuable advantages to users. However, the auditory feedback style for the sensory and hybrid modes could benefit from further refinement (e.g. simplification, distance values added to objects, distances communicated via 'beep rate') to align better with user expectations, which should enhance their ratings on these scales.

AI-Sight App progression (semantic-mode, hybrid-mode) in response to beta-testing feedback

The qualitative and quantitative feedback from beta-testers illustrates several paths forward to improving the practicality and usability of each mode (sensory, semantic, hybrid). This can be categorized into information changes, audio changes, and object recognition changes:

- *Information changes*: Adding distance information to the semantic and hybrid-modes was highly requested, both in terms of ascribing a specific distance to specific objects in the scene, but also as a safety backup in case specific objects are not recognized by the AI. We are currently adding a series of new information processing steps to create this in an intuitive manner for users as well as have it run reliably and still in real-time.
- *Audio changes*: Beta-testers expressed universal acclaim for the verbal feedback in terms of clarity and brevity. This will be further enhanced via adding distance descriptions and giving App users' rapid updates on any changes in their spatial relationship to the object. The role of abstract audio (sensory, hybrid) was seen as being in a better place to provide additional context, for example, tones providing peripheral awareness of other objects/options that they may want to select or shift their focus to for the verbal feedback. Users generally preferred distances to objects to be communicated by changes in timing of beeps, like a typical parking sensor. Our new changes in how distance and object information is processed by AI-Sight will allow this new method of audio feedback to be implemented for the sensory and hybrid modes. We will revise our live-modes to have verbal feedback focused on providing a high level of detail in the center, and instead have tones focus more on peripheral information and hazard alerts.
- *More relevant object recognition*: There was also a request for additional object categories to be recognized, and for these objects to be more relevant to specific tasks that users would like to accomplish, rather than generic objects within the environment. Many of these tasks relate to instrumental activities of daily living (IADLs) such as food preparation, navigating bathrooms, clothing identification as well as navigating shops, public parks, and homes. We are progressing this on two fronts: training AI to recognize new objects relating to IADLs; and building AI-Sight's features on new smartphone AI models with additional object categories (e.g. YOLOv8). We discuss some of the progress with this below.

AI-Training and distance-estimation papers

Our research paper on training YOLOv5 to detect 35 objects of interest to blind and low vision users was published as a PubMed indexed paper at IEEE EMBC, and presented at their conference in Sydney, Australia in late July 2023 [2]. A second research paper exploring how five different approaches to measuring distances using the same platform (iPhone 13 pro) perform in terms of accuracy in the center and periphery of the image alongside a range of usability metrics (CPU usage, battery usage, field-of-view) was submitted to IEEE OJEMB, and the revised manuscript was resubmitted following peer review with a minor revision recommendation.

In addition, we have worked on training and tuning the object recognition capabilities of relevant AI-models towards recognizing new object categories of interest to persons who are blind or low vision. Each path is with a view to eventually implementing these models into a current or future version of AI-Sight. This work includes:

- Object Segmentation (DeepLabV3) – Here we have worked on training from scratch a DeepLabV3 model (used in the current AI-Sight) towards more categories of indoor objects. This uses the ADE20k dataset, which is a comprehensive set of object segmentation training data that can be sub-divided into specific situations (e.g. living room, bathroom). We have initially evaluated these models in terms of their accuracy, comparing general indoor models against bathroom, kitchen, and bedroom-specific models. We also explored the roles of different training image sizes and image darkness. We are currently preparing a manuscript on this evaluation.
- Object Detection (YOLOv8) – Here we explored how we can improve recently available AI models that recognize (but cannot segment yet on mobile devices) 80 object categories with transfer learning techniques. Here we added four new object categories relevant for indoor navigation (door, door handle, etc) and track how they could be effectively added to a pre-trained model and how they affected recognition of the pre-existing categories. We are currently preparing a manuscript on this assessment.

While DeepLabV3 training is done with a view to implementing these models into the current AI-Sight, making use of YOLOv8 training will involve implementing into a future version of AI-Sight that is built on this AI-model.

Overall, the development of SoundSight's sensory-mode and initial versions of AI-Sight's semantic and hybrid modes (both live, and button-selected) provided our participants with experience of a wide range of approaches to conveying 'visual' information through sound for tasks at home. Their feedback provided a series of directions to revise the mobile technologies to be more usable and practical for the next batch of beta-testers. Some revisions, such as adding distance information to specific objects and new forms of audio-feedback will be ready for the second batch of beta-testers, while others are longer-term goals, with training and integrating new AI-models focused on their requested objects, that we aim

to be ready for the third batch of beta-testers. Each approach adds specific forms of valuable information for the users, and so balancing the overall soundscape to provide this information in unison but also in a clear manner remains a top priority. Here beta-testing feedback has been invaluable in identifying where this balance lies.

References
[1] Commère, L., & Rouat, J. (2023). Evaluation of Short-Range Depth Sonifications for Visual-to-Auditory Sensory Substitution. *IEEE Transactions on Human-Machine Systems*, 53(3), 479-489.
[2] Sankarnarayanan, T., Paciorkowski, L., Parikh, K., Hamilton-Fletcher, G., Feng, C., Sheng, D., ... & Chan, K. C. (2023, July). Training AI to Recognize Objects of Interest to the Blind and Low Vision Community. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1-4). IEEE.

What opportunities for training and professional development has the project provided?

Nothing to Report

How were the results disseminated to communities of interest?

Our work training AI to recognize objects of interest to the blind and low vision community was published in IEEE EMBC, with our distance-estimation paper resubmitted to IEEE OJEMB after minor revision recommendation. This research and the Apps have been presented during undergraduate and postgraduate seminars to students from NYU's Center for Data Science. This is with a view to showing how computer-vision approaches can benefit a wide range of assistive technologies, and has led to a range of relevant computer-vision and assistive technology projects at NYU Langone Health.

We are currently revising the AI-Sight App to provide both the requested information and preferred auditory feedback methods from the initial batch of beta-testers. Once completed, we will recruit the next batch of beta-testers to assess these improvements during at-home use. During this process we will also work to implement our newly trained AI-models that recognize more object types into the AI-Sight, as well as start work on a YOLOv8 version of the AI-Sight. We will also add new 'sensory-level' feedback to the AI-Sight (rather than the SoundSight) so all modes exist in one App. The final batch of beta-testers will assess these AI changes alongside any further requested revisions from the second batch of testers. We will then write a technical evaluation paper for the AI-Sight detailing its design and user feedback.

4. IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

The technology developed in this project is unique as it combines the most advanced and accurate 3D sensing available on smartphones (LiDAR) with an accurate form of object segmentation (DeepLabV3). This process turns complex environments into simplified/prioritized images that show the type, number, size, and 3D location of objects. This can be conveyed intuitively via verbal feedback, or via 'sensory substitution' which turns the visual image into abstract audio soundscapes. These approaches complement one another, with the verbal feedback being more accessible to the object identities, and sensory substitution being more accessible to the object features. This study shows how persons with blindness or low vision can use each approach and how they interact with one another as they solve daily tasks. Understanding ideal combinations of information and feedback will further increase the usability and practicality of these assistive technologies and solve long-standing issues with sensory substitution by making them easier to learn independently.

What was the impact on other disciplines?

Nothing to Report.

What was the impact on technology transfer?

Nothing to Report.

What was the impact on society beyond science and technology?

Nothing to Report.

5. CHANGES/PROBLEMS:

Changes in approach and reasons for change

Nothing to Report.

Actual or anticipated problems or delays and actions or plans to resolve them

Previously, integrating depth information into AI-Sight also introduced significant usability issues (e.g. smaller field-of-view). As a result, batch 1 of beta-testers did not have access to depth in their version of the AI-Sight and instead experienced distances through our SoundSight App. The beta-testers' preferred approaches to audio feedback required us to not only add distances but have a high level of integration between the depth and object-segmentation images (i.e. look up all distance values related to each recognized object and assign it the closest value). We have since integrated LiDAR distance estimation without the prior usability issues and co-registering the depth and object-segmentation images to assign objects their closest distances in real-time. The high level of detail requested by users has taken additional time to create and translate into their desired audio feedback in preparation for batch 2 of testers.

Changes that had a significant impact on expenditures

N/A

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Significant changes in use or care of human subjects

No significant changes are noted. Recruitment is set to continue for batches 2 and 3 as the technology reaches the desired milestones.

Significant changes in use or care of vertebrate animals

N/A

Significant changes in use of biohazards and/or select agents

N/A

6. PRODUCTS:

- **Publications, conference papers, and presentations**

Journal publications.

One submitted and under review:
Hamilton-Fletcher, G., Liu, M., Sheng, D., Feng, C., Hudson, T., Rizzo, J.-R. & Chan, K. C. (resubmitted and under review after minor revision recommendation). Accuracy and Usability of Smartphone-based Distance Estimation Approaches for Visual Assistive Technology Development. Submitted to: *IEEE Open Journal of Engineering in Medicine and Biology.*

Books or other non-periodical, one-time publications.

Nothing to Report

Other publications, conference papers and presentations.

One published and presented at IEEE EMBC 2023:
Sankarnarayanan, T., Paciorkowski, L., Parikh, K., Hamilton-Fletcher, G., Feng, C., Sheng, D., Hudson T.E., Rizzo, J.R. & Chan, K. C. (2023, July). Training AI to Recognize Objects of Interest to the Blind and Low Vision Community. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1-4). IEEE*. doi: 10.1109/EMBC40787.2023.10340454.

- **Website(s) or other Internet site(s)**

Nothing to Report

- **Technologies or techniques**

The development of 'hybrid-mode' is novel as it integrates two technologies (sensory substitution and semantic segmentation) to produce a new way for persons who are blind to experience AI-simplified images. These technologies will be shared via future publications and a public release (end of grant period) and could serve as useful new rehabilitative and scientific tools. We have also been training new AI-models focused on the needs of persons who are blind which further enhance the usability and practicality of these approaches.

- **Inventions, patent applications, and/or licenses**

N/A

- **Other Products**

Nothing additional to report.

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name: *Kevin C. Chan, Ph.D.*
Project Role: *Principal Investigator*
Researcher Identifier (e.g. ORCID ID): *ORCID ID: 0000-0003-4012-7084*
Contribution to Project: *no change*

Name: *John-Ross Rizzo, M.D.*
Project Role: *Co-Investigator*
Researcher Identifier (e.g. ORCID ID): *ORCID ID: 0000-0002-4084-0085*
Contribution to Project: *no change*

Name: *Todd E. Hudson, Ph.D.*
Project Role: *Co-Investigator*
Researcher Identifier (e.g. ORCID ID): *ORCID ID: 0000-0003-4506-2670*
Contribution to Project: *no change*

Name: *Giles Hamilton-Fletcher, Ph.D.*
Project Role: *Post-Doc researcher*
Researcher Identifier (e.g. ORCID ID): *ORCID ID: 0000-0001-5903-4334*
Contribution to Project: *no change*

Name: *Dean Sheng*
Project Role: *Graduate Student / Research Assistant*
Researcher Identifier (e.g. ORCID ID): *N/A*
Contribution to Project: *no change*

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

(1) PI's previously active grants have closed. They include:

(i) NIH R01 EY028125 grant titled "Glaucoma neuroimaging in humans and experimental animal models "

(ii) ARVO Foundation for Eye Research grant titled "Magnetic resonance imaging of healthy and glaucomatous eyes".

(2) PI's pending NIH grant RF1 NS126102-01S1 as Co-I is now active. The project title is "High-resolution bidirectional optical-acoustic mesoscopic neural interface for image-guided neuromodulation in behaving animals"

(3) Co-investigator JR Rizzo's pending NIH grant 4R33EY033689-03 grant is now active. The project title is "VIS4ION-Thailand (Visually Impaired Smart Service System for Spatial Intelligence and Onboard Navigation)".

(4) Co-investigators JR Rizzo's and Todd Hudson's previously active NSF grant has closed. The project title is "NSF Convergence Accelerator Track H: Smart Wearables for Expanding Workplace Access for People with Blindness and Low Vision" .

There are no scientific or budget overlaps for the above grants with the current DoD grant.

What other organizations were involved as partners?

.

Nothing to Report

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS: ***N/A***

QUAD CHARTS: *Updated and submitted accordingly to eBRAP.*

9. APPENDICES: N/A