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 2022

TYPE OF REPORT: Annual Report

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

DISTRIBUTION STATEMENT: Approved for Public Release; Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
OCTOBER *2022*	Annual Report	1SEPT2021 - 31AUG2022

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
VR200130

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. These Apps convey information gathered by sensors (color, distance, heat) and artificial intelligence (object recognition) through both spoken verbal feedback and 'musical' audio that intuitively conveys the locations, visual properties, and identities of objects in the environment. Our research purpose is to produce new Apps that are tailored to blind end users to increase visual information accessibility, enhance daily functionality, and facilitate new interactions of interest to them. In terms of scope, this 2-year project focuses on the development of novel technologies in the first year, with at-home beta-testing by fully blind subjects in the second year. The technology development focuses on new iPhone sensors (LiDAR range-finding, plug-in thermal cameras) and their support for state-of-the-art object recognition techniques that run in real-time, locally on iPhone. During year 1, we found that DeepLabV3, which accurately segments object shapes from live visual images, provides new interaction possibilities. Here the location, shape, size, and identity of recognized objects in a scene can be rapidly presented to users through 'musical' feedback. This represents scenes at the 'semantic-level' (objects). This goes beyond prior technologies that operate at the 'sensory-level' (e.g. brightness, distances, heat) to provide a more intuitive understanding of the environment that remains stable across variable conditions. Furthermore, since object identity is known, we provide optional verbal feedback that tells users each object's name and describes its location in the image. This provides live user support and training within the App. Building on this, we are preparing for user testing in year 2. Here blind end users will beta-test our Apps that convey information at various levels (e.g. sensory, semantic). Through interviews and questionnaires, our end user feedback will help us further refine our technologies to better address user interests and enhance its impact on day-to-day function.

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** USAMRDC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified	Unclassified	Unclassified	Unclassified	**19**	

TABLE OF CONTENTS

1. **INTRODUCTION:** *Narrative that briefly (one paragraph) describes the subject, purpose and scope of the research.*

'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 names), as well as interact 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:** *Provide a brief list of keywords (limit to 20 words).*

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

3. **ACCOMPLISHMENTS:** *The PI is reminded that the recipient organization is required to obtain prior written approval from the awarding agency grants official whenever there are significant changes in the project or its direction.*

What were the major goals of the project?
List the major goals of the project as stated in the approved SOW. If the application listed milestones/target dates for important activities or phases of the project, identify these dates and show actual completion dates or the percentage of completion.

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. motion, color, location – Major Task 1)
(2) 'semantic information' (e.g. "table", "door" – Major Task 2)
(3) a sensory and semantic 'hybrid mode' through auditory feedback (spatialized tones & 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

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, heat) from smartphone cameras / sensors via auditory feedback. 80% of features complete but ready for beta-testing.

Given a starting month of September 2021, the timeline lists Major Task 1 (sensory-only mode) as having a target date of January 2022. Currently, we have multiple modes running which span various sensors (camera, distance, thermal) and variety of audio modes (e.g. tones, 'hot' sounds). To aid in ease of use, we will be building multiple versions of the sensory-mode App, dedicated to specific sensors for deployment via Apple's TestFlight beta-testing distribution App – e.g. a 'Thermal SoundSight' and a 'Distance+Color SoundSight'. This means that users will be given 'optimal' versions of each sensor-audio combination (as determined through internal testing) as individual App instances on their phone. User feedback in the future may indicate further desired changes in their settings (**subtask 1, 90%**).

Subtask 2 by contrast looks at extracting/eliminating continuous planes such as floors and walls from the image. The use of DeepLabV3 in our semantic and hybrid modes automatically eliminate these planes which gives us the ability to ask beta-testing subjects whether the elimination of these planes is beneficial to their understanding of the environment through comparing the sensory and hybrid modes. The original method of using Apple's ARKit for plane-detection was deprioritized due to the above alternative ways of gathering the same information, and time taken to create the new hybrid mode, which is detailed later. ARKit's foundation is still implemented within the SoundSight App, and we are currently looking into implementing ARKit into our 'semantic-mode' and 'hybrid-mode' Apps to expand our future potential technical development options during year 2 (**subtask 2, 70%**).

Major Task 2 (Months 4-8): Develop and implement computer-vision technologies in the smartphone App that recognize everyday objects and delivers as verbalized semantic names. 80% of features complete, but ready for beta-testing.

Major task 2 (semantic-only mode aka "Computer Vision (CV) App") has a timeline of Feb-May 2022. The CV App uses DeepLabV3 to identify and segment recognized objects from the environment in real-time (~8 frames per second). The CV App tracks the location and identity of multiple objects simultaneously within the image with 20 types of objects being recognized (**subtasks 1 & 5, 100%**). The object's horizontal position is indicated through verbal feedback descriptors after the object name is spoken ("bottle, far left"; "car, middle"). The object's vertical position is denoted through the speakers' vocal pitch (low = low-pitched 'male' voice; high = high-pitched 'female' voice), which also helps separate different objects as each object and its location is spoken by a unique voice (**subtask 3, 80%**). While horizontal and vertical positions are given, we currently lack providing distance information due to narrow field-of-view restrictions imposed by Apple when using our main way of working with depth (Apple's "AVDepthData"), we are exploring alternative distance measurements using Apple's ARKit method (**subtask 2, 50%**). Object identity and location updates in real-time (8 frames per second) and upon a double tap of the screen, the user has this information verbalized for them (**subtask 4, 100%**). Objects are cycled through from leftmost-to-rightmost to provide an easy-to-understand, predictable structure for the user (**subtask 5, 100%**). Due to the additional difficulties with integrating distance information into the semantic-level App, this has been prioritized over the thermal integration, which can still be accessed in the sensory-mode App (**subtask 6, 0%**). As the 'hybrid-mode' App (Major Task 3) is built on top of the semantic-mode App, we have the ability to switch between semantic and hybrid representations in real-time simply through the user either doing a 'long-press' or 'double tap' on the screen, representations can be presented individually, or queued up one after the other (**subtask 7, 80%**), the sensory-only mode remains a separate App, supporting different sensors, that can be switched to. In addition to the stated subtasks, we have also integrated object 'thresholding' to avoid false-positive recognitions (e.g. a couple of pixels recognized as a 'person' when there is no person in the image), such as ignoring specific object categories if their size is under 2% of the image. This produces more stable and robust list curation for verbal feedback.

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

Major task 3 (hybrid-mode App) has a timeline of June-Sept 2022. Since the original proposal and setting of subtasks, recent advanced object-recognition techniques such as DeepLabV3 have opened new doors for how hybrid approaches can be done to improve practicality and ease-of-presentation, and the work below reflects this. The 'CV App' contains both the semantic-mode and hybrid-mode functionalities, using DeepLabV3 to

segment the image to provide sensory features such as location/size/shape for recognized objects (**subtask 1, 100%**). Users have control over the feedback of information presented, and at any moment are able to hear the names and locations of objects verbally (semantic-mode) by double tapping the screen; or hear the location, shape, size, distribution, and identity of objects through abstract 'musical' feedback that 'paints' the image using sound (hybrid-mode - **see figure 1**) by holding a 'long press' on the screen (**subtask 2, 90%**). New audio specifically made for hybrid mode has been created. Here, different sounds are played for different objects using naturalistic metaphors (e.g. "bottle" = glass clink; "car" = horn), with variations in pitch and panning indicating how these objects are distributed across the vertical and horizontal axes in the image. The hybrid-mode scan-through takes ~4 seconds, and if the user double-taps on the screen during this scan, the object names are read out afterward for the same image, meaning both modes co-exist sequentially (**subtask 3, 90%**). We are exploring additional ways interleaving sensory and semantic feedback. Finally, we are also implementing constant feedback during 'live-mode' to provide general contextual awareness.

We believe the advances made here for the hybrid mode are suitable for writing for technical publication after further refinements, as we are not aware of any sensory substitution devices that operate at the 'object level' in an image to provide object identity, shape and location. This is important because the core features provided to users are 'sensory-invariant' – in that objects are just as easily extracted and presented to users, even if the object and background are the same color, which in general massively increases the difficulty when using a traditional sensory substitution device. We are planning to start writing for publication in a technical journal after the first group is recruited for end-user testing, which would have us finish reaching **milestone 3**.

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

Major Task 4 (Preparation: Months 7-9; Recruitment: Months 10-21): Documents required for the public beta-testing have previously been approved by both the USAMRDC Human Research Protection Office (**Subtask 1, 100%**) and NYU Langone Health Institutional Review Board (**Subtask 2, 100%**) by April 2022. Now that the final 'Hybrid-mode' in App development is suitable for beta-testing, we have revisited the protocol, interviews, and questionnaires to determine suitability or additional questionnaires of interest. Subject enrollment is at the preparation stages and involved collating our lists of potential volunteers with blindness who have asked us to contact them when these studies open up to enrollment (**Subtask 3**, **10%**). This will open the path for full subject recruitment after the final software and hardware preparation is complete.

What was accomplished under these goals?

For this reporting period describe: 1) major activities; 2) specific objectives; 3) significant results or key outcomes, including major findings, developments, or conclusions (both positive and negative); and/or 4) other achievements. Include a discussion of stated goals not met. Description shall include pertinent data and graphs in sufficient detail to explain any significant results achieved. A succinct description of the methodology used shall be provided. As the project progresses to completion, the emphasis in reporting in this section should shift from reporting activities to reporting accomplishments.

Please read below our goal reporting for this period.

Since the last reporting period 3 months ago, our work has focused on the technical development of the 'Hybrid-mode' App, hardware preparation, accessibility testing, and revisiting the interviews/questionnaires to ensure their suitability for current use since their original approval.

In terms of technical development within the last reporting period we have focused on creating the 'Hybrid-mode' App (**Specific aim 1, Major task 3**). This involves sampling the DeepLabV3 object-level image at 10 by 10-pixel grid locations, with this information selecting constituent sounds (object type = instrument-style; vertical position = pitch; horizontal position = left-right panning and time). These constituent sounds together play over time creating a musical soundscape that is informative to object identity, size, and location (**see figure 1**). This development has been built on the 'CV App' foundation which already produces the 'semantic-mode' feedback (verbal feedback for object identity and location). This has the benefit of automatically integrating development from the semantic- and hybrid-modes into one App. This allows the sensory features of objects (shape, size, location) to be represented through 'musical soundscapes' like our sensory-level only sensory substitution device (SSD) (the SoundSight) as well as verbal feedback from semantic-mode.

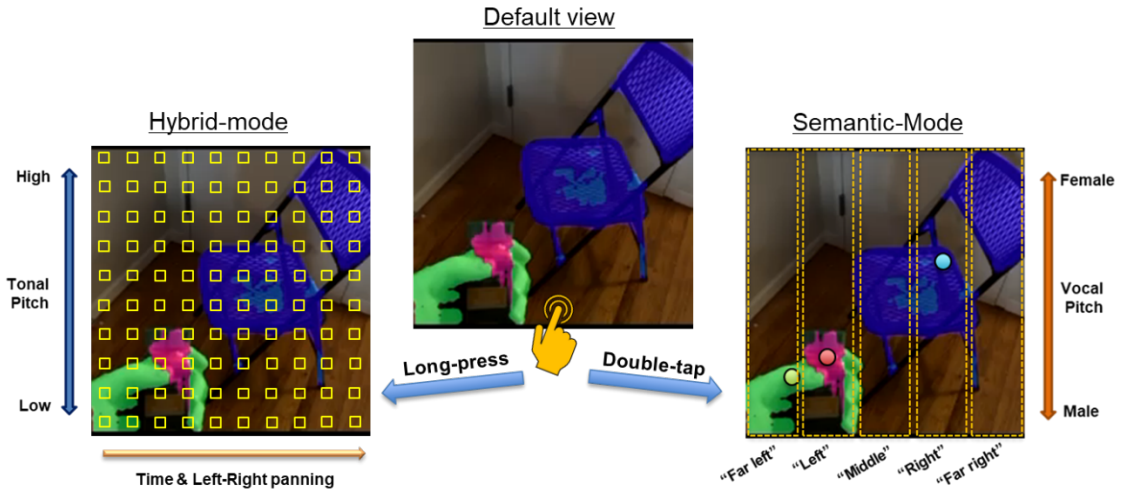

Figure 1. Illustration of how the CV App currently works. Upon launching the App, the user is presented with colored overlays of objects recognized by the system (middle image). If the user does a 'long-press' of the screen, Hybrid-mode plays. In Hybrid-mode (left image) each location (yellow boxes – purely illustrative) is checked for the presence of a recognized object. Objects (20-types) are categorized into Groups (8-types), which determines the sound file selected to play for that location. Columns are played left to right over time. For this image, the user will hear a low-pitched Choir ("person") on the left, then a low-pitched glass 'clink' ("bottle"), and finally tones ("chair") that ascend from middle to high-pitched. If the user double-taps the screen, 'semantic-mode' plays. Here objects a verbally read-out, from leftmost-to-rightmost, with object center-points determining the horizontal description and the vocal pitch of the voice saying the object.

The hybrid-mode representation selects 100-pixel locations in a 10 by 10 grid format from DeepLabV3's segmentation map. The segmentation map is a representation of the image, but each 'pixel' value relates to

what recognized object is within that location (0 = nothing; 1 = aeroplane; 2 = bicycle…). For each column of 10 pixels, this information – the pixel's location and object value – is used in a decision tree to select the appropriate sound file (instrument-timbre, pitch, and volume) and is loaded into memory in advance of being played. These sound files are then 'released' (or played) in a rapid succession with an intentional but slight (15ms) difference in starting time between each sound file and the next in line. For a full vertical column this gives the impression of running a finger across multiple piano keys, which helps listeners identify the number of pixel locations with objects as well as any gaps (as opposed to playing 10 sounds that start simultaneously, which our prior research indicates is a source of difficulty for listeners) [1]. During the playback of these 10 audio files, the next column's audio files are selected, loaded into memory, and prepared for playback. This column's audio files can start playing before the prior columns have finished, which speeds the scan through time as well as helping produce a constantly active soundscape. This process repeats for all 10 columns, from left-to-right, with spatial panning also changing from leftmost to rightmost in alignment with the column positions. Finally, an optional drumbeat is also included, the purpose of this is to ensure that some audio is playing during the scan-through even if no objects are detected. This confirms to the listener that the scan is progressing and allows the listener to track the scan's lateral position. The total scan takes approximately 4 seconds. This completes **Subtask 1**.

Since the semantic and hybrid-modes exist within the same App, the user is given gestural controls to select which representation they wish to listen to. Here a 'double-tap' produces a semantic representation of the image (e.g. "Bottle, far left… Person, middle… TV, right"), while a 'long-press' (1 second hold) activates a hybrid-mode scan of the image. This also includes the ability to 'queue-up' representations. As such during a 'hybrid-mode' scan, if the user double-taps the screen, 'semantic-level' verbal feedback of that same image will occur afterwards. The ability for the user to listen to representations separately or combine in the order they wish allows the user to direct their own learning and feedback in natural environments. This completes **Subtask 2**.

The audio that is used for the 'hybrid-mode' needs to do several things:
- intuitively represent a variety of objects (to avoid rote memorization)
- variations of this audio can communicate pixel location (e.g. different pitches)
- be aurally distinct from the verbal feedback (so both can coexist)
- be aesthetically pleasing (to address a common criticism of SSDs)

Since DeepLabV3 recognizes 20 types of objects, that could mean potentially 20 different qualities of audio (e.g. instrument-types) may are required. This might be too many abstract concepts for users to easily remember, so to reduce complexity we have started by representing 8 object categories through 8 types of audio:

- **Danger** (bus, car, motorbike, train, aeroplane, bicycle) = "trumpet horn"
- **Person** (person) = "choir"
- **Pet** (cat, dog) = "meow"
- **Bird** (bird) = "bird chirp"
- **Bottle** (bottle) = "glass bottle clink"
- **TV** (TV) = shakuhachi (sounds like "TV static noise")
- **Furniture** (chair, sofa, table) = "soft tone"
- **Miscellaneous** (cow, boat, horse, plant, sheep) = "harp"

The quality of sound is chosen based on a 'natural metaphor' system which we explored in our paper on the SoundSight system [2]. Here sounds that could be reasonably assumed to uniquely emanate from or belong to an object, are used to aurally represent these objects. For example, the 'danger' category contains multiple transport-related objects, many of which can utilize horns to announce their presence in day-to-day life. As such, one 'natural metaphor' sound is the "trumpet horn" which sounds like a car horn but is aesthetically pleasing and musical (and so satisfies our other requirements). The benefit of this approach is that the mappings are not arbitrary, and are informed by shared cultural knowledge and day-to-day living experience. As a result, listeners do not need as much rote memorization from training and can instead use their intuitions to arrive at/near the object category. The downside of our current approach is that multiple objects can be clustered under the same category, and so further clarification from verbal feedback is the only way to

disentangle these when only using the App. In the future we will seek create additional audio to further subdivide object categories (e.g. musical "woofs" for 'dog' from the 'pet' category).

The audio files used for object categories are MIDI-generated timbres (beepbox.co), and each has 10 different pitch variations. The pitches span 10 notes on the pentatonic scale. This scale is used in other SSDs like the EyeMusic [3] due to the lack of 'clashing' notes, so any combination of notes is aesthetically pleasing. These notes span A4 (440Hz) to C3 (130Hz). Higher pitches relate to higher pixel locations in line with the design decisions of many prior SSDs, this mapping between pitch and height is psychologically intuitive according to cross-modal correspondences research [4]. The creation of object-specific audio for hybrid-mode, its pitch variations, and its ability to precede or follow semantic-level verbal feedback within the App completes **Subtask 3**.

Since both the semantic and hybrid modes require active selection by the user, our internal discussions raised the point that the user still needs to know 'when is a good time to explore the scene further'. To address this point, we have started implementing continuously active but simpler scene feedback. This produces verbal feedback when a new object enters the scene, but does so using a more conservative threshold then semantic-mode (e.g. an object might need to take up 15% of the image, rather than 2%, in order to be spoken). This conservative thresholding avoids false-positives and over-burdening the list of objects that takes time to be read out (longer list read-outs can become "out-of-date" by the time the names are spoken).

Hardware and deployment preparation

The iPhone 13pro devices for App deployment have been updated through NYU Langone MCIT security to run within an NYU Langone Health approved "container" environment, so that their usage can be remotely monitored and controlled by NYU Langone MCIT security. In terms of hardware, we are waiting for MCIT's suggestions for protective cases and will order the remaining thermal cams during the next purchase order. Deployment for the App will be done through Apple's TestPilot App, in preparation for this we have an Apple developer license and are securing the relevant data for deployment on these specific phones (Unique Device ID numbers, serial number, profile name etc). This will allow remote 3-month testing periods for the Apps on the iPhones.

User study preparation

With the Apps now in a state ready for beta-testing, monthly discussions have revisited the protocol, interview, and questionnaires to ensure the protocol written in Dec 2021 is still the best approach given our progress in terms of software, hardware, and user feedback methods. In terms of devices, our protocol seeks to test sensory, semantic, and hybrid representations of the environment from the end users' perspective. Since writing the protocol we have found two complementary Apps (Super LiDAR, Lidar Sense), both of which use a single point of distance and turn that into tactile feedback, with one of these Apps also using computer-vision for 3 object-classes (door, person, seat). We are considering including these on subject's iPhones as a complementary way to gather additional information regarding tactile feedback (which we are currently not exploring) and some additional object-classes from DeepLabV3.

In terms of the interview and protocol, we are looking into adjusting the language of the App from SoundSight to "provided Apps" since not all features are within one App. We are also going to add questions that ask which object categories they are most interested in. The purpose of this is to help guide future DeepLabV3 A.I.-training efforts that seek to recognize objects of interest to persons with blindness. To aid A.I.-training we are seeking advice from the Center for Data Science at NYU to examine how this could progress. In order to streamline the final interview further we are looking into offloading some of the questions to other standardized questionnaires in the literature. The two questionnaires we are considering are the "System Usability Scale" (SUS) and the "NASA-Task Load Index" (NASA-TLX). These two questionnaires are commonly used to evaluate Apps/devices/products, with SUS asking subjects to rate their level of agreement/disagreement to 10 statements on (e.g. "I felt very confident using the system"), while NASA-TLX asks subjects to rate their experiences on scales exploring mental/physical/time-demands as well as performance, effort, and frustration. The use of these may enhance comparisons with other technologies evaluated using SUS and NASA-TLX.

Issues/delays with specific App features

Since software development shifted to principally building on the 'CV App' foundation (which contains DeepLabV3 semantic segmentation), we have worked to include additional sensory features such as depth/distance, which could prove helpful in adding to the semantic and hybrid-modes. However, as described in the prior quarterly report, our existing methods for implementing depth (Apple's AvDepthData) radically reduced the camera's field-of-view (from ~60 to ~15 degrees) which massively reduces environmental context while increasing the precision required of the user to point the camera 'correctly' (Figure 2). We explored the use of the 'selfie-cam' 3D LiDAR as well, but despite its better field-of-view, requires the user to hold the phone backwards, which besides being counter-intuitive, also makes App interactions more difficult (Figure 2). The most promising approach for implementing distances without significant downsides is ARKit (Apple's augmented reality development environment). This requires transferring the code base of the CV App from A**V**CaptureSession to A**R**Session, and ensuring that the original features work before integrating the depth information that ARSession has access to. Overall, while depth/distance information exists in the SoundSight (as well as distance-to-tactile stimulation LiDAR Apps) the unexpected pitfalls in using the same methods for the CV App has slowed this specific avenue of development. We still intend to push this area forward, particularly using ARKit, during year 2.

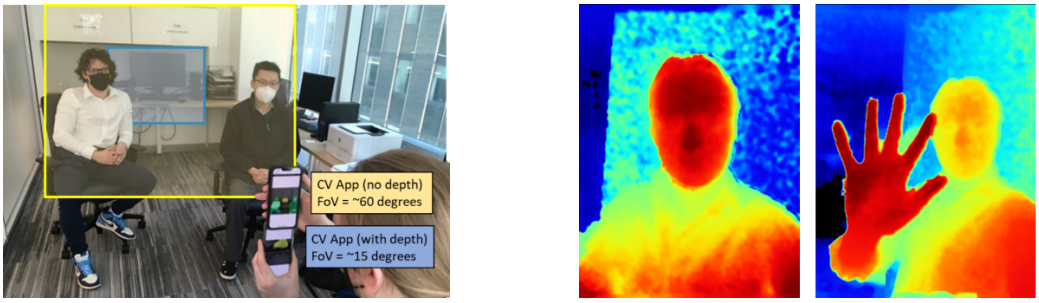

Figure 2 – Leftmost image: Semantic-mode CV App running on two iPhones, segmenting people (green) and the monitor (yellow). The top iPhone does not have depth enabled but natively provides a wider field of view (~60°, yellow box) than when depth is enabled (~15°, blue box). As a result of the smaller field-of-view crucial objects are missed. Future directions seek to work around this. Rightmost image: Depth information from 'TrueDepth' which is a front-facing 'selfie' LiDAR system on iPhone. Since this approach uses the front-facing camera, for use as a sensory-tool to explore the world, the user would have to reverse the iPhone to face away from them, making it unwieldy, and user interactions difficult.

[1] Hamilton-Fletcher, G., & Chan, K. C. (2021, November). Auditory Scene Analysis Principles Improve Image Reconstruction Abilities of Novice Vision-to-Audio Sensory Substitution Users. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 5868- 5871). IEEE.

[2] Hamilton-Fletcher, G., Alvarez, J., Obrist, M., & Ward, J. (2022). SoundSight: a mobile sensory substitution device that sonifies colour, distance, and temperature. Journal on Multimodal User Interfaces, 16(1), 107-123. doi: 10.1007/s12193-021-00376-w

[3] Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., & Amedi, A. (2014). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. Restorative neurology and neuroscience, 32(2), 247-257.

[4] Spence, C. (2011). Crossmodal correspondences: A tutorial review. Attention, Perception, & Psychophysics, 73(4), 971-995.

What opportunities for training and professional development has the project provided?
If the project was not intended to provide training and professional development opportunities or there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe opportunities for training and professional development provided to anyone who worked on the project or anyone who was involved in the activities supported by the project. "Training" activities are those in which individuals with advanced professional skills and experience assist others in attaining greater proficiency. Training activities may include, for example, courses or one-on-one work with a mentor. "Professional development" activities result in increased knowledge or skill in one's area of expertise and may include workshops, conferences, seminars, study groups, and individual study. Include participation in conferences, workshops, and seminars not listed under major activities.

Nothing to Report

How were the results disseminated to communities of interest?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the results were disseminated to communities of interest. Include any outreach activities that were undertaken to reach members of communities who are not usually aware of these project activities, for the purpose of enhancing public understanding and increasing interest in learning and careers in science, technology, and the humanities.

Advancements in this research and the CV App have been presented as introductory presentations to computer-vision orientated students (Data Science seminar at NYU, both at undergraduate and postgraduate levels). This is with a view to showing how computer-vision approaches can benefit assistive technologies, but also how object recognition methods can be trained to focus on objects of interest to persons with blindness.

Describe briefly what you plan to do during the next reporting period to accomplish the goals and objectives.

We are currently recruiting volunteers from blind subjects to participate in the beta testing in Year 2. We will also continue to refine and finalize the 3 modes of the App, finalize the questionnaires for interview, and interact with students and colleagues when presenting our work at both undergraduate and postgraduate levels for projects relating to training and assessing object detection, semantic segmentation, and depth estimation methods. Insights gained from this effort should help to better tune our assistive technology systems towards our specific use cases (e.g. blind user indoors).

4. **IMPACT:** D*escribe distinctive contributions, major accomplishments, innovations, successes, or any change in practice or behavior that has come about as a result of the project relative to:*

What was the impact on the development of the principal discipline(s) of the project?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how findings, results, techniques that were developed or extended, or other products from the project made an impact or are likely to make an impact on the base of knowledge, theory, and research in the principal disciplinary field(s) of the project. Summarize using language that an intelligent lay audience can understand (Scientific American style).

The development of the hybrid mode is unique within the assistive technology field. It combines two lines of research: (1) **Sensory substitution devices (SSDs)**: Converting visual images into sound for blind users, and (2) **Semantic segmentation**: "cutting out" objects from the environment in real-time on smartphones. These have a unique synergy as SSDs are effective at communicating basic shapes through audio to users, and 'semantic segmentation' distills the complex natural environment down to simple object shapes (e.g. a rectangle TV). This creates a unique level of information transfer, as object identity (e.g. "TV") can be used to inform the specific audio feedback (e.g. TV static sounds), or constrain which parts of the image are conveyed to users (e.g. colors within recognized objects). This could address major shortcomings of current SSDs and provide new rehabilitation/ scientific tools.

What was the impact on other disciplines?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the findings, results, or techniques that were developed or improved, or other products from the project made an impact or are likely to make an impact on other disciplines.

Nothing to Report.

What was the impact on technology transfer?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe ways in which the project made an impact, or is likely to make an impact, on commercial technology or public use, including:

- *transfer of results to entities in government or industry;*
- *instances where the research has led to the initiation of a start-up company; or*
- *adoption of new practices.*

Nothing to Report.

What was the impact on society beyond science and technology?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how results from the project made an impact, or are likely to make an impact, beyond the bounds of science, engineering, and the academic world on areas such as:

- *improving public knowledge, attitudes, skills, and abilities;*
- *changing behavior, practices, decision making, policies (including regulatory policies), or social actions; or*
- *improving social, economic, civic, or environmental conditions.*

Nothing to Report.

5. CHANGES/PROBLEMS: *The PD/PI is reminded that the recipient organization is required to obtain prior written approval from the awarding agency grants official whenever there are significant changes in the project or its direction. If not previously reported in writing, provide the following additional information or state, "Nothing to Report," if applicable:*

Problems: As detailed above, while some development features have progressed as expected or faster, one aspect that provided unexpected difficulty was integrating depth/distance information into our second App (CV App) without significant downsides (smaller field-of-view; non-optimal camera direction). While we look to resolve this through integrating ARKit, this has adversely affected the completion of relevant subtasks for the CV App.

Changes: Monthly meeting feedback has also suggested the additional inclusion of some questionnaires and Apps. Questionnaires could streamline feedback with the System Usability Scale (SUS) to assess how users felt interacting with the system, and NASA-TLX to assess how demanding it is to use. Similar questions can be removed from the 3-month interview. Apps that communicate distance-to-tactile feedback are also discussed for inclusion into the protocol. This allows additional feedback on relevant areas that we did not plan to explore in our App development.

Actual or anticipated problems or delays and actions or plans to resolve them
Describe problems or delays encountered during the reporting period and actions or plans to resolve them.

Solving the problems stated above have required more development time than expected. As such, we prioritized development paths that we were more confident would lead to tangible results. While we wish to integrate ARKit depth into the CV App, this will have to occur in year 2, with feedback on distances coming from the SoundSight (and other Apps). Fortunately, gathering feedback on the CV App before and after depth-integration may provide insights on its value. The integration of further questionnaires and Apps requires a resubmission of the protocol for IRB approval, which may delay subject recruitment while we wait for approval to be obtained.

Changes that had a significant impact on expenditures

Describe changes during the reporting period that may have had a significant impact on expenditures, for example, delays in hiring staff or favorable developments that enable meeting objectives at less cost than anticipated.

N/A

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

Describe significant deviations, unexpected outcomes, or changes in approved protocols for the use or care of human subjects, vertebrate animals, biohazards, and/or select agents during the reporting period. If required, were these changes approved by the applicable institution committee (or equivalent) and reported to the agency? Also specify the applicable Institutional Review Board/Institutional Animal Care and Use Committee approval dates.

Significant changes in use or care of human subjects

No significant change is noted as human subject components will occur only in Year 2 but not Year 1.

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:** *List any products resulting from the project during the reporting period. If there is nothing to report under a particular item, state "Nothing to Report."*

- **Publications, conference papers, and presentations**
  Report only the major publication(s) resulting from the work under this award.

  Journal publications. *List peer-reviewed articles or papers appearing in scientific, technical, or professional journals. Identify for each publication: Author(s); title; journal; volume: year; page numbers; status of publication (published; accepted, awaiting publication; submitted, under review; other); acknowledgement of federal support (yes/no).*

Nothing to Report

  Books or other non-periodical, one-time publications. *Report any book, monograph, dissertation, abstract, or the like published as or in a separate publication, rather than a periodical or series. Include any significant publication in the proceedings of a one-time conference or in the report of a one-time study, commission, or the like. Identify for each one-time publication: author(s); title; editor; title of collection, if applicable; bibliographic information; year; type of publication (e.g., book, thesis or dissertation); status of publication (published; accepted, awaiting publication; submitted, under review; other); acknowledgement of federal support (yes/no).*

Nothing to Report

  Other publications, conference papers and presentations. *Identify any other publications, conference papers and/or presentations not reported above. Specify the status of the publication as noted above. List presentations made during the last year (international, national, local societies, military meetings, etc.). Use an asterisk (*) if presentation produced a manuscript.*

Presentations were given at the NYU Langone Health Department of Ophthalmology presentation day, and for NYU Center for Data Science in order to receive feedback and to seek collaborations to benefit future development.

- **Website(s) or other Internet site(s)**
  List the URL for any Internet site(s) that disseminates the results of the research activities. A short description of each site should be provided. It is not necessary to include the publications already specified above in this section.

Nothing to Report

- **Technologies or techniques**
  Identify technologies or techniques that resulted from the research activities. Describe the technologies or techniques were shared.

The development of the 'semantic-mode' is a new technology here but shares similarities with other verbal-feedback Apps. The development of 'hybrid-mode' is genuinely novel in integrating two separate technologies together (sensory substitution and semantic segmentation). These technologies are due to be shared in the future via publications and public release (end of grant period) and could serve as useful new rehabilitative/scientific tools.

- **Inventions, patent applications, and/or licenses**
  Identify inventions, patent applications with date, and/or licenses that have resulted from the research. Submission of this information as part of an interim research performance progress report is not a substitute for any other invention reporting required under the terms and conditions of an award.

N/A

- **Other Products**
  Identify any other reportable outcomes that were developed under this project. Reportable outcomes are defined as a research result that is or relates to a product, scientific advance, or research tool that makes a meaningful contribution toward the understanding, prevention, diagnosis, prognosis, treatment and /or rehabilitation of a disease, injury or condition, or to improve the quality of life. Examples include:
  - *data or databases;*
  - *physical collections;*
  - *audio or video products;*
  - *software;*
  - *models;*
  - *educational aids or curricula;*
  - *instruments or equipment;*
  - *research material (e.g., Germplasm; cell lines, DNA probes, animal models);*
  - *clinical interventions;*
  - *new business creation; and*
  - *other.*

Nothing additional to report.

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Provide the following information for: (1) PDs/PIs; and (2) each person who has worked at least one person month per year on the project during the reporting period, regardless of the source of compensation (a person month equals approximately 160 hours of effort). If information is unchanged from a previous submission, provide the name only and indicate "no change".

Name: Kevin C. Chan, Ph.D.
Project Role: Principal Investigator
Researcher Identifier (e.g. ORCID ID): ORCID ID: 0000-0003-4012-7084
Nearest person month worked: 0.9 /year
Contribution to Project: Project overview, technology overview, human subject protocol/ recruitment / assessments.

Name: John-Ross Rizzo, M.D.
Project Role: Co-Investigator
Researcher Identifier (e.g. ORCID ID): ORCID ID: 0000-0002-4084-0085
Nearest person month worked: 0.6 /year
Contribution to Project: 'Semantic-mode' development, computer-vision, sensors, assessments for human subject research.

Name: Todd E. Hudson, Ph.D.
Project Role: Co-Investigator
Researcher Identifier (e.g. ORCID ID): ORCID ID: 0000-0003-4506-2670
Nearest person month worked: 0.6 /year
Contribution to Project: 'Semantic-mode' development, computer-vision.

Name: Giles Hamilton-Fletcher, Ph.D.
Project Role: Post-Doc researcher
Researcher Identifier (e.g. ORCID ID): ORCID ID: 0000-0001-5903-4334
Nearest person month worked: 6.0 / year
Contribution to Project: Technology overview, 'sensory-mode' development, sensors, audio output, human subject recruitment / assessments.

Name: Dean Sheng
Project Role: Graduate Student / Research Assistant
Researcher Identifier (e.g. ORCID ID): N/A
Nearest person month worked: 6.84 / year
Contribution to Project: 'Semantic-mode' development, computer-vision, audio output.

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

If the active support has changed for the PD/PI(s) or senior/key personnel, then describe what the change has been. Changes may occur, for example, if a previously active grant has closed and/or if a previously pending grant is now active. Annotate this information so it is clear what has changed from the previous submission. Submission of other support information is not necessary for pending changes or for changes in the level of effort for active support reported previously. The awarding agency may require prior written approval if a change in active other support significantly impacts the effort on the project that is the subject of the project report.

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

(i) Research to Prevent Blindness/Stavros Niarchos Foundation grant titled "In vivo, Multi-modal Characterization of Neurodegeneration and Neuroprotection of the Visual System in a Novel Experimental Glaucoma Model."

(ii) BrightFocus Foundation grant titled "The Role of Brain Waste Clearance System in Glaucoma".

(iii) Feldstein Medical Foundation titled "Role of Insulin Resistance in the Eye and Brain"

(2) PI's pending NIH grant R01-EY033353-01 as Co-I is now active. The project title is "Neurotoxicity of Reactive Astrocyte-Secreted Lipids in Neurodegenerative Disease"

(3) Co-investigator JR Rizzo's pending NIH grant 1R21EY033689-01A1 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 and Crown Castle/CATT grants have closed. The project titles are "Collaborative Research: Future expert work in the age of "black box", data- intensive, and algorithmically augmented healthcare" and "Semantic Visual-Inertial SLAM: Advanced Wearables Reconstructing Scenes Without infrastructure".

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

What other organizations were involved as partners?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe partner organizations – academic institutions, other nonprofits, industrial or commercial firms, state or local governments, schools or school systems, or other organizations (foreign or domestic) – that were involved with the project. Partner organizations may have provided financial or in-kind support, supplied facilities or equipment, collaborated in the research, exchanged personnel, or otherwise contributed.

Provide the following information for each partnership:
Organization Name:
Location of Organization: (if foreign location list country)
Partner's contribution to the project (identify one or more)

- *Financial support;*
- *In-kind support (e.g., partner makes software, computers, equipment, etc., available to project staff);*
- *Facilities (e.g., project staff use the partner's facilities for project activities);*
- *Collaboration (e.g., partner's staff work with project staff on the project);*
- *Personnel exchanges (e.g., project staff and/or partner's staff use each other's facilities, work at each other's site); and*
- *Other.*

Nothing to Report

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS: *For collaborative awards, independent reports are required from BOTH the Initiating Principal Investigator (PI) and the Collaborating/Partnering PI. A duplicative report is acceptable; however, tasks shall be clearly marked with the responsible PI and research site. A report shall be submitted to* ***https://ebrap.org/eBRAP/public/index.htm*** *for each unique award.*

QUAD CHARTS: *If applicable, the Quad Chart (available on https://www.usamraa.army.mil/Pages/Resources.aspx) should be updated and submitted with attachments.*

Updated and submitted accordingly to eBRAP.

9. APPENDICES:

Attach all appendices that contain information that supplements, clarifies or supports the text. Examples include original copies of journal articles, reprints of manuscripts and abstracts, a curriculum vitae, patent applications, study questionnaires, and surveys, etc.