Visually Impaired Users can Locate and Grasp Objects Under the Guidance of Computer Vision and Non-Visual Feedback

Nii Mante, James D. Weiland, *Fellow, IEEE*

Abstract— The purpose of this study was to assess the ability of blind individuals to reach and grasp for objects, under the guidance of auditory (verbal) or vibrotactile cues controlled by real time computer vision algorithms. For these experiments, we created the Object Localization and Tracking System (OLTS). The OLTS utilized a head mounted wide-angle (diagonal 92° degrees) monocular camera, a central processing unit and two types of physical feedback: auditory bone conduction headphones or cranially positioned vibration motors. A computer vision algorithm, the Context Tracker, processed live video to track objects in front of the visually impaired subject. Physical feedback was then generated based on the object position. The feedback guided the user to move the camera until the desired object was in the central region of the camera, defined by an angle of the camera field of view. The central region was varied between 3.9 and 39.6 degrees. Experiments consisted of localizing and grasping for an object based on feedback provided. On average, subjects were able to locate the correct object within 20 seconds. For auditory feedback, using a central angle of 7.8° led to poor performance compared to the other angles. Performance using vibrotactile feedback worsened when using a central angle of 3.9°. No consistent performance trends were evident based on age of blindness onset.

I. Introduction

The National Academies of Sciences, Engineering and Medicine recently issued a population health imperative to promote eye health and to reduce vision impairment.[1] Visually impaired individuals are required to do simple daily tasks in a manner different than their sighted peers. Reaching and grasping for an object of interest is an example of a task that is difficult for the visually impaired.

There have been a few systems that aim to solve the problem of finding and grasping objects (object localization). The vOICe, GroZi, Brainport, and OrCam are examples of such systems. The vOICe and Brainport use sensory substitution, that is, use of a different sensory modality to sense and interpret information typically processed by vision. The vOICe converts video, from a head mounted camera, to a "soundscape", i.e. a complex sound that corresponds to the video. The user must interpret this soundscape in order to understand what is in the view of the camera.[2] The Brainport translates video from a head mounted camera into electrical stimulation patterns on the surface of the tongue.[3] Specifically, this system relies on the subject's ability to process the raw electrical patterns to determine shape, size, location and motion of objects. GroZi and OrCam rely on computer vision to interpret some information and this understanding is relayed to the user. GroZi, a prototype handheld grocery shopping assistant for the visually impaired, offers object recognition, via computer vision algorithms, for finding items and haptic/speech feedback for grasping the items. It also utilizes scene text recognition for overhead aisle signs. In this system, the camera is located on the handheld device itself. The OrCam offers optical character recognition via a camera that attaches to eyeglasses.[4] It gives voice feedback via a bone conduction earphone. However, it's input modality is different; the OrCam reads text only in the region indicated by the user (with hand gestures), which implies that the user must know where the object is to begin with. For people with little or no vision, knowing the location of the desired object remains a significant problem.

We have reported previously on the system design and initial testing of a wearable visual aid with a head mounted camera that performs object recognition and tracking tasks and uses auditory cues for guidance.[5] Subsequent discussions with potential users revealed that auditory cueing may be problematic in situations when blind people use their natural hearing to detect auditory cues from the external environment. Vibrotactile cues are an alternative method that would not occupy hearing. In the experiments described in this manuscript, we compare the efficacy of auditory and vibrotactile cuing as feedback during an object acquisition task. The experimental system for tracking/feedback evaluation is referred to as the Object Localization and Tracking System (OLTS).

II. Methods

A. System Overview

The OLTS acquires video using a head-mounted wide field-of-view (FOV) camera (92°, 480x640 pixels, 30 frames/sec, NTSC format), tracks an object's position by means of computer vision algorithms, and provides feedback to the user on how to move the camera to center the object in the camera FOV. Once the object is centered, the user can rely on proprioception to grasp the object in front of them, based on head-pointing. To guide the user's head movements, the

* This research and development project/program/initiative was conducted by the University of Southern California and is made possible by a cooperative agreement that was awarded and administered by the U.S. Army Medical Research & Materiel Command (USAMRMC) and the Telemedicine & Advanced Technology Research Center (TATRC), at Fort Detrick, MD under Contract Number W81XWH-10-2-0076

N. Mante was with the Dept. of Biomedical Engineering, University of Southern California.

J. D. Weiland is with the Dept. of Biomedical Engineering, University of Michigan, Ann Arbor, MI USA (weiland@umich.edu).

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

system generates auditory feedback via bone conduction headphones, or vibrotactile feedback via pancake motors positioned on a glasses frame. The bone conduction headphones used in these experiments were Gamechanger LLC *Audiobone* 1.0 headphones. The pancake motors used were Beam Bristlebot Robot 14 mm x 3.6 mm pancake micro motors (applied voltage 1.3V - 3.0V). Tracking and feedback are performed in real-time.

A tracking algorithm is used to follow, frame-to-frame, an object as it moves in the camera's FOV while the subject moves their head. For these experiments the object is initially selected by the operator. The algorithms were run on an Apple Macbook Pro (4 GB RAM, 2.4 GHz i5 processor). The tracking algorithm was developed by Dinh and Medioni.[6] This algorithm, the Context Tracker, utilizes contextual information as a fundamental guiding principle that enhances object tracking, and is built upon the principles of the Tracking-Learning-Detection (TLD) tracker. [7]

The final stage of the OLTS system is the feedback algorithm. The algorithm uses the object position from the tracking algorithm to give auditory or vibrotactile feedback to the visually impaired subject. The acquisition of the object's position and feedback generation is done in real time. Due to the fact that our camera is monocular, 3D depth information is not readily available. Thus, we ignore the depth element of object localization and worry only about the 2D position of the object within the field of view. Once the feedback algorithm has the object's 2D position, as provided by the vision algorithms, the position is passed to our *Sensory Map*.

The *Sensory Map* translates the 2D position of the object to a discrete 'code.' (Figure 1) There are nine discrete codes, and each corresponds to a region within the visual field of the camera. The values of the 'code' are as follows: *UL, UR, U, L, C, R, DL, D, DR*. Depending upon the code provided, a spoken word or phrase is played to the subject or one or two motors are activated. Specifically, once the centroid of the object's bounding box overlaps with any portion of a region, then the feedback for that region is generated. For example, if the centroid of the object resides within the lower right portion of the cameras FOV, the *DR* code is passed to the speech synthesis and the words "Down and Right" are spoken to the subject. Similarly, the fourth motor, *M4*, would vibrate when given the *DR* code. The subjects are instructed on the appropriate response for each cue.

Four vibration motors are used to allow us to generate 8 different combinations of vibrations. For example, vibrating motors 2 and 4 together represents the "right" command from the sensory map. Vibrating only motor 3 represents the "down and left" command for the sensory map. Using two motors would only allow for three possible patterns (left, right or both vibrating). Using more would be unnecessary based on the sensory map design. If the system detected 3D position, more motors could be used to dictate a depth pattern or vibration intensity could be modulated to encode depth.

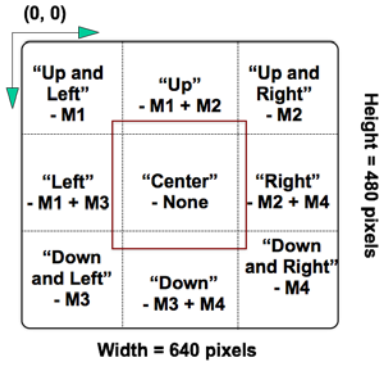

Figure 1. Sensory Map for the auditory (in quotes) and vibrotactile feedback mechanisms. The grid represents the camera's field-of-view (FOV). An object's position in 3D space can be mapped to the 2D FOV above. Once the object is mapped to this FOV, it will fall into one of the 9 grid locations. Depending on the location of the object within the grid, the computer will generate the corresponding word/vibration for the subject. Auditory feedback is comprised of computer-generated speech, and vibrotactile feedback is achieved by triggering of 4 vibration motors in different combinations.

Figure 2. Experimental Setup. (Top) The subject is seated, and the camera is mounted on glasses worn by the subject. (Bottom) The vision algorithm tracks the object of interest (object highlighted by green bounding box). The user is instructed to grasp the object once they receive the center command from the computer. In the case of auditory, a "center" command is a speech-synthesized voice saying "Center." The "center" command for vibrotactile is *no vibration.* In this case, the center region was 31.2°.

2.2 Experimental Protocol

A localization experiment required the subject to use the OLTS to reach and grasp for a desired grocery store item out of a set of 3-5 items in front of them (Figure 2). The objects' straight-line distance (SLD) from the subject ranged from 0.15 m to 0.66 m, based on the subject's reach. The subject's goal was to centralize the desired object in the camera FOV

using auditory/vibrotactile feedback from the OLTS, and then reach and grasp for the object.

At the beginning of each localization experiment, a set of objects were placed in front of the subject. The researcher could see a camera feed of the head mounted camera. The only requirement in this step was for the desired object to be within the FOV. Distractor items were always immediately adjacent to desired object (Figure 2). The researcher would then draw a bounding box around the desired object, and start the autonomous tracking and feedback stage. Several types of data were recorded, including the time to completion, number of reaches per task, object tracking path, the video stream during each task, and the commands provided to the subject.

After training the subjects on the use of OLTS, 10 trials were done for each of the five central visual angles (7.8, 15.6, 23.4, 31.2, 39) for a total of 50 trials per subject for each feedback modality. 13 visually impaired subjects were tested using both the vibrotactile (n = 12) and auditory (n = 13) feedback modalities. One subject was not available to return for testing vibrotactile feedback. The testing phase normally lasted for ~1.5 hours. To avoid biases related to the order of testing to the visual angle, the subjects were split into 3 groups, each with a different order of visual angle. In a second round of experiments, we extended the visual angle to 3.9 degrees for the vibrotactile, since we did not see diminished performance at 7.8 degrees (see Results).

All testing was approved by the University of Southern California Institutional Review Board, and performed at the Braille Institute in Los Angeles, California. Subjects were read the informed consent form, and then signed the form to enroll in the study. Background medical information was obtained on their eye condition both from their ophthalmologist and from a questionnaire, under HIPAA regulations.

III. Results

The four measurements were time to first grasp, time to completion, number of attempts and first grasp success rate. The time to first grasp indicates the amount of time it took the subject to grasp any object. Time to completion indicates the amount of time it took for the subject to successfully grasp the *correct* object. The number of attempts indicates how many "reaches" it took for the subject to grasp the *correct* object. Lastly, the first grasp success rate indicates how often the subject grabbed the *correct* on the first attempt. Time to completion is shown in graphical form, the remaining data is described in the text.

For auditory feedback, a one-way ANOVA analysis showed that there was a significant difference for the time to completion ($F_{4, 644} = 29.1$, $p < .05$), and time to first grasp ($F_{4, 644} = 42.2$, $p < .05$) between the five central visual angles. Post-hoc tests showed that the center angle of 7.8° (smallest angle) was the reason for the statistical difference. A center angle of 7.8° yielded slower completion times and first grasp times in comparison to other angles. There were no statistically significant differences amongst the remaining angles (15.6°, 23.4°, 31.2° and 39°) for any other measures.

Our results from the first set of experiments showed no statistically significant difference between the five visual angles for any of the measures when using vibrotactile feedback (Figure 3). A lower performance threshold was found at 7.8 for auditory feedback. To determine if a lower performance threshold existed in vibrotactile feedback, additional object localization experiments were done in 5 of the original 12 subjects. Three angles (3.9°, 15.6°, 39°) were tested. Two of the angles (15.6°, 39°) were retested along with the new angle (3.9°). Including two angles from the original testing was done to compare performance to the first set of experiments. These additional results showed a statistically significant difference in time to 1st grasp ($F_{2, 147} = 7.6$, $p < .05$) and completion time (Figure 4) ($F_{2, 147} = 4.7$, $p < .05$) between 3.9°, and the remaining two angles. There was no statistically significant difference for amount of reaches and 1st grasp success rate between the different visual angles ($p > .05$). Additionally, the old and new data from these 5 subjects for angles 15.6° and 39° were compared. Analysis showed that there was no statistically significant difference in performance from these different days ($p > .05$).

Subjects scanned more to find objects using vibrotactile cues. We counted the number of pixels in the desired object path when tracked. The number was generally greater when using vibrotactile cues (Figure 5), indicating that the subjects searched more.

Subject Comments

In order to gain insight on the system design, subjects were questioned after using both feedback mechanisms. There was no definitive trend that suggests more people would prefer sound feedback versus vibrotactile feedback. However, most subjects stated that vibration is better suited in louder environments, since the auditory feedback would be difficult to hear. Subjects liked the direct and specific nature of the auditory feedback. Their impression was that the auditory feedback directly "tells you what to do", whereas with vibrotactile feedback "you have to feel it out." One common statement among patients was the possibility of switching between feedback modalities for different situations via a physical switch.

Vibrotactile vs. Auditory

For each angle, the time to completion was compared between the vibrotactile and auditory via one-way ANOVA. For angles 7.8° and 23.4° there was a statistically significant difference in performance between vibrotactile and auditory feedback ($F_{1, 247,} = 10.6$ $p < .05$), ($F_{1, 248,} = 10.6$ $p < .05$). For angles 15.6°, 31.2° and 39° there was no significant difference in performance between vibrotactile and auditory feedback ($p > .05$). Thus, although a statistical difference was noted for 2 angles, performance was not clearly better for auditory or vibrotactile feedback.

One-way ANOVA was done for each angle between early and late blind. Specifically, the time to completion for early vs. late blind subjects was compared for auditory and vibrotactile feedback. For vibrotactile feedback, the results show that there was a statistically significant difference in performance between early (n = 5) and late blind (n = 7) for angles 23.4°

($F_{1,118} = 7.1$, $p < .05$) and 31.2° ($F_{1,118} = 11.4$, $p < .05$). In both of those angles, the early blind subjects performed better. For auditory feedback, the results show that there was a statistically significant difference in performance between early (n = 4) and late blind (n = 9) for the angle 15.6° ($F_{1,127} = 18.4$, $p < .05$). In this angle for auditory, the late blind performed significantly better. The remaining angles for auditory showed no significant difference in performance between early and late blind ($p > .05$). Overall, no clear trends were evident when comparing early vs. late blind subjects.

IV. Conclusion

We have presented a wearable, experimental system that allows the visually impaired to reach and grasp for objects. The system is built upon computer vision tracking techniques in conjunction with custom feedback algorithms and modalities. We have presented quantitative measurements of object localization tasks. Our analysis shows auditory feedback led to worse performance (slower time to completion) at a central visual angle of 7.8°. Additionally, vibrotactile performance worsened at an angle of 3.9°. Vibrational signals may allow a smaller central window since the user may react faster to such signals, whereas a spoken instruction takes time to deliver and understand. But this explanation require more experimentation. The results suggest that an additional step, using computer vision techniques to positively identify the object, is needed. This final step will act as the confirmation step, and effectively close the loop in our wearable system. This prototype can serve as the basis for a mobile system that uses wearable cameras and processors, like those used for augmented reality.[8]

Acknowledgment

The authors thank the Los Angeles Braille Institute for assisting our experiments by allowing us to use their facilities and work with their students.

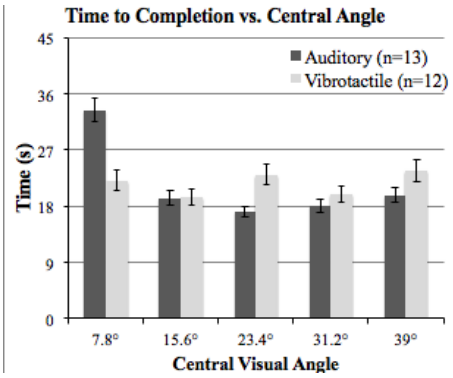

Figure 3. Subjects were tested using 5 central angle sizes. Auditory cues took more time at the smallest angle, whereas vibrotactile cues took the same team regardless of angle.

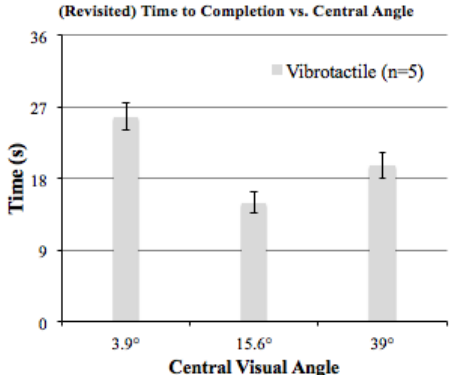

Figure 4. Retesting of vibrotactile feedback to find lower performance. A central angle of 3.9° increased completion time.

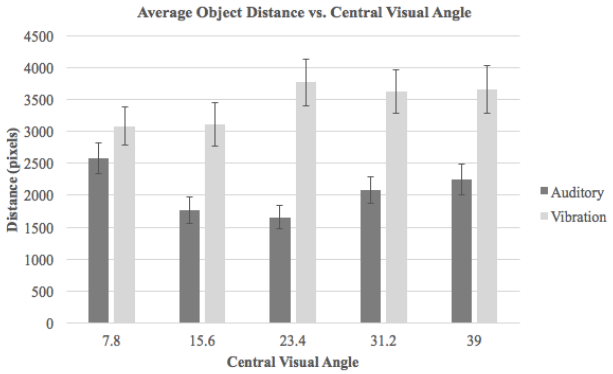

Figure 5. The distance in pixels traveled as a function of angle and type of cue. Except for the smallest angle, subjects searched more for the central region when using vibrotactile cues, meaning they scanned back-and-forth less when using auditory cues.

References

[1] Teutsch SM MM, Woodbury RB, Welp A Making Eye Health a Population Health Imperative: Vision for Tomorrow. . In. *The National Academies Press*. Washington DC: The National Academies of Sciences, Engineering, and Medicine.; 2016.

[2] Ward J, Meijer P. Visual experiences in the blind induced by an auditory sensory substitution device. *Conscious Cogn.* 2010;19(1):492-500.

[3] Bach-y-Rita P, S WK. Sensory substitution and the human-machine interface. *Trends Cogn Sci.* 2003;7(12):541-546.

[4] Waisbourd M, Osama A, Siam L, Moster MR, Hark LA, Katz LJ. The Impact of a Novel Artificial Vision Device (OrCam) on the Quality of Life of Patients with End-Stage Glaucoma. Paper presented at: ARVO 20152015; Denver, CO.

[5] Thakoor K, Mante N, Zhang C, et al. A system for assisting the visually impaired in localization and grasp of desired objects. Paper presented at: European Conference on Computer Vision2014.

[6] Dinh TB, Vo N, Medioni G. Context tracker: Exploring supporters and distracters in unconstrained environments. Paper presented at: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on2011.

[7] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence.* 2012;34(7):1409-1422.

[8] Trese MG, Khan NW, Branham K, Conroy EB, Moroi SE. Expansion of severely constricted visual field using Google Glass. *Ophthalmic Surgery, Lasers and Imaging Retina.* 2016;47(5):486-489.