A System for Assisting the Visually Impaired in Localization and Grasp of Desired Objects

Kaveri Thakoor, Nii Mante, Christian Siagian, James Weiland, Laurent Itti, and Gerard Medioni

University of Southern California
Los Angeles, California, USA
{thakoor, mante, itti, medioni}@usc.edu, christian.g.siagian@gmail.com, jweiland@med.usc.edu

Abstract. A prototype wearable visual aid for helping visually impaired people find desired objects in their environment is described. The system is comprised of a head-worn camera to capture the scene, an Android phone interface to specify a desired object, and an attention-biasing-enhanced object recognition algorithm to identify three most likely object candidate regions, select the best-matching one, and pass its location to an object tracking algorithm. The object is tracked as the user's head moves, and auditory feedback is provided to help the user maintain the object in the field of view, enabling easy reach and grasp. The implementation and integration of the system leading to testing of the working prototype with visually-impaired subjects at the Braille Institute in Los Angeles (demonstration in the accompanying video) is described. Results indicate that this system has clear potential to help visually-impaired users in achieving near-real-time object localization and grasp.

Keywords: object recognition, attention, tracking, localization, grasp, auditory feedback, visually impaired

1 Introduction

The World Health Organization estimates that there are 285 million visually impaired people in the world [1]. Studies by Nau et al. [2] have concluded that locating items is one of the prime tasks for which visually impaired persons continue to depend on sighted helpers. This paper describes a prototype wearable visual aid, currently under development, which provides near-real time object recognition, localization, tracking, and guidance cues to help a visually impaired user to point his or her head towards a desired object, allowing for easy reach and grasp of the target object. The system consists of an Android-phone-based command module, which allows the user to query for specific items of interest via finger-touch. A two-stage computer vision algorithm next localizes and recognizes the desired item: recognition is achieved via SURF (Speeded Up Robust Features: robust local feature detector developed by Bay et al. [26]), further enhanced by an attention-biasing algorithm developed at the University of Southern California (Attention Biased Speeded Up Robust Features, AB-SURF). This

algorithm [5] takes inspiration from the human cognitive system's ability to bias attention to characteristics relevant to the desired object and uses these features to locate the three most likely candidate regions in the scene where the object may be present. Next, high-performing Speeded Up Robust Features (SURF) [26] are extracted to find which one of these three possible regions best matches images of the desired object contained in the trained database; location information regarding this best region is passed to an object tracker [32]. The tracker maintains position information of the object as the user's head moves, while auditory feedback cues are provided to help the user center the object in the camera field of view for easy access. Visually impaired subjects were able to use the system to grasp desired objects within 12 to 13 seconds from the time the algorithm recognized the object, demonstrating the system's value for improving the independence and quality of life for visually impaired people.

2 Related Work: System Overview and Context

Object recognition devices for the visually impaired can be categorized into two main groups: (1) wearable, camera-based systems which process the incoming scene [6, 7] and (2) smart-phone or finger-pointing based aids that recognize items within the field of view of the phone camera or region defined by the finger-point [8, 9, 10, 11]. Of the first variety, Bjorkman et al. [6] have implemented a system which utilizes two pairs of stereo cameras, taking inspiration from the human visual process: one pair of cameras is for saccading to relevant object regions, and the other pair is for foveating (or focusing) for finer grained recognition within this relevant region. Furthermore, they utilize hue-based saliency and SIFT (Scale Invariant Feature Transform) [25] based features along with color histograms for object recognition; they also harness depth information using the stereo cameras to determine object size and filter objects based on foreground or background location. Although this approach shows good accuracy on household objects and functions quickly, its bulkiness makes it impractical for use by individuals, and there exists no means of providing guidance or feedback to a user once an object is found. On the other hand, in the work of Bigham et al. [8], the user captures an initial picture of the scene using a smartphone (lightweight and easy to use, though relies on the user's ability to frame a well-defined scene); the image is then crowd-sourced: expert sighted human annotators view the images on a website and provide back detailed segmentation and object identity information, allowing for object localization for the visually-impaired user (via auditory feedback) but relying on external annotators to obtain that goal. The OrCam [12] introduced in June of 2013 has combined both of these approaches with an eye-level camera attached to the user's glasses and basic recognition capability (of traffic lights, signs, and some objects); however, it also relies on the user having sufficient vision to point his or her finger to relevant items of interest in the surroundings. The novelty of the innovation described here lies in its combination of the best of both of the above approaches while also providing a closed loop system (that can aid even

those who are completely blind, as no localization on the part of the user is required). The system utilizes only a single camera with lightweight, portable algorithms (all implemented on a single MacBook Pro I5, 2.4 GHz laptop computer), that attend, recognize, localize, track, and provide feedback to the user in near real-time from the time the user provides a tactile request for an object on an Android phone. All system components can be seen in the figure below.

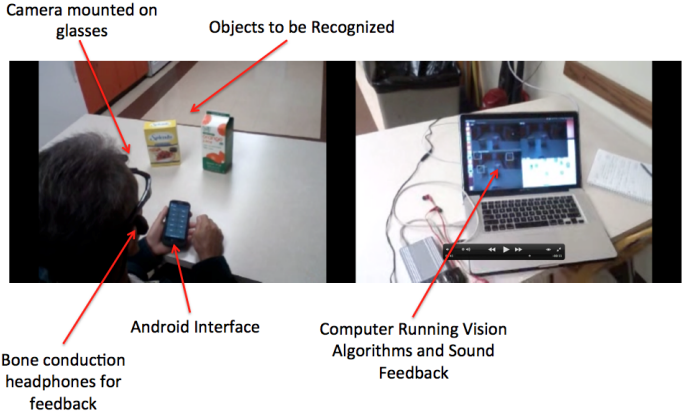

Fig. 1. All components of system shown as they are typically set during an experiment with a visually impaired subject.

3 Implementation Details

3.1 Android User Command Interface

This section describes the implementation of the front-end user interface that the visually impaired user employs to provide input to the system as to which item he or she is interested in locating. A User Datagram Protocol (UDP) server-client model was chosen to allow communication between a Java-based tactile interface application on an Android phone [27], which acts as the client, and a Neuromorphic Robotics Toolkit (NRT) C++ module on the MacBook Pro laptop mentioned earlier, which acts as the server and houses all computer vision algorithms for the system (NRT is the modular toolkit developed at the University of Southern California [34] that is used for system integration throughout this paper). The Android device transmits messages in the form of strings (the name of the user's object query, e.g. "SPLENDA"), while the server module receives this string information and transforms it into a form that can be understood and processed by the rest of the object recognition chain (also built within separate NRT modules, discussed in more detail in Section 4).

3.2 Object Recognition

Alternate Recognition Techniques Although several smart-phone based applications and electronic travel aids exist (including [9, 10, 12]) that provide recognition and navigation assistance to blind and low vision users, they require the user to select a 'frame' or region as input to the system. This is a real challenge for visually impaired users [2, 3, 4]. Thus, a computer vision system that can localize a desired region or object within a scene and provide recognition of the contents of that region is invaluable. Taking inspiration from the human cognitive system, the guided visual search model introduced by Jeremy Wolfe [13, 14] explains that humans locate items of interest in their surroundings by focusing attention on 'targets' (key features relevant to the current query item) and paying less attention to 'distractors'. While other object recognition algorithms exist that rely on modulating attention [15], recognition is restricted to certain dataset types (cars, faces). Specifically to help the visually impaired in their grocery shopping, Winlock et al. [16] have designed a system to recognize several grocery store items using a combination of mosaicing, SURF descriptors, and Bayesian statistics; they have also established certain objects that are easier to recognize than others. While achieving near real-time recognition, their system does not contain an explicit closed-loop feedback component for localization after recognition. We accomplish this by harnessing attention biasing to fixate on regions most relevant given a user's query for a specific item among objects from a trained database, followed by SURF (Speeded Up Robust Features) for object recognition in the chosen regions. As another alternative, sensory substitution devices (SSDs) are a unique class of aids for the blind that use a modality other than vision to convey visual information to the user. An example of an SSD is the vOICe system. The vOICe SSD represents the camera-captured scene via an auditory soundscape. The vOICe conversion program [17] transforms a scene into auditory information ('soundscapes') based on three rules: the vertical axis (i.e., elevation of the object) is represented by frequency, the horizontal axis by time and stereo panning, and the brightness of the image is encoded by loudness. Although these conversion rules appear relatively simple, explicit and quite extensive training is required to learn how to interpret even simple shapes. Learning to use the vOICe SSD requires months of training before independent utility can be achieved [18].

While SSDs, electronic travel aids, and smart phone applications may provide some form of scene understanding, they suffer from various disadvantages: excessive reliance on human neuroplasticity and learning before a blind person can use them productively, inability to locate the relevant regions where desired objects may be present in a scene, and potentially uncomfortable latency times before a recognition response is provided to a user. Our work, Attention-Biased Speeded Up Robust Features (AB-SURF), provides both localization and recognition and is optimized for integration into a closed loop system which also tracks the recognized object and provides real-time feedback, allowing the user to independently query, seek, and reach out to grasp a desired object in the surroundings.

Traditional computer vision object recognition models rely on sliding-window detectors (i.e. application of object recognition algorithms on overlapping subimages of the scene, tessellating the whole scene, a computationally intensive and time consuming process) [19, 20]. To enhance efficiency, we have replaced this technique by an attention-based mechanism that narrows down regions of interest by biasing the search toward statistical features consistent with the query object. Attention mechanisms are broken into three different types: (1) those stemming from bottom-up cues [21, 22], e.g. inherently conspicuous regions, (2) those influenced by top-down (task-driven) motivation, e.g. knowledge of the target item [24], and (3) those resulting from a blend of bottom-up and top down forces [23, 24]. The bottom-up saliency maps are computed via the method of Itti and Koch [24]. Specifically, in our top down approach, we use 30 hues and 6 saturations within the color dimension, 12 intensities within the luminance dimension, and 8 orientations. Overall, this narrows the space over which recognition must compute, yielding a five-fold speedup as compared to SURF performance alone on 640 x 480-pixel images. In addition, we achieve object recognition accuracy with AB-SURF in simple cluttered scenes comparable to the recognition accuracy of SURF on single, isolated objects (more quantitative results discussed in next section).

Neurally-Inspired Object Recognition: Attention-Biased Speeded Up Robust Features (AB-SURF) Here we describe implementation and performance of an algorithm to localize as well as recognize objects contained in regions of interest, named Attention Biased Speeded Up Robust Features (AB-SURF). AB-SURF functions by computing a biased saliency map; unlike a bottom-up saliency map [24], which is generated by determining the most inherently salient regions in a scene by extracting features across channels of hue and orientation and weighing all features equally, a biased saliency map is generated by heavily weighing those characteristics most consistent with the item of interest, allowing for attention to be focused to regions relevant for a given query even if they are not necessarily the most conspicuous regions at first glance. Once these regions are extracted, SURF-based [26] object recognition is used to evaluate these top three regions to determine the best match given the query, outputting only the top result as the recognized object.

Complete analysis of attention-biased SURF object recognition was carried out on 382 10-object images and 655 5-object images in [5]; True Positive Rates (TPRs: number of instances when the object was correctly recognized as present out of the total number of tested images) are plotted for each object below. All five objects present in the 5-object images exhibit true positive recognition rates of greater than 80%, with three of the five objects having recognition rates of greater than 99%. Overall, the 5-object case yields excellent results. Furthermore, attention biasing significantly reduces the computation required for recognition by eliminating the need for a brute force sliding window approach to locate a desired object in the image. The sliding window approach (at 50% window overlap) would require at least 40 subwindows (or many more, if the sliding window

overlap is greater than 50% to improve performance) to be recognized per frame, instead of just the 3 subwindows selected by attention-biasing, requiring 7 to 8 seconds to process.

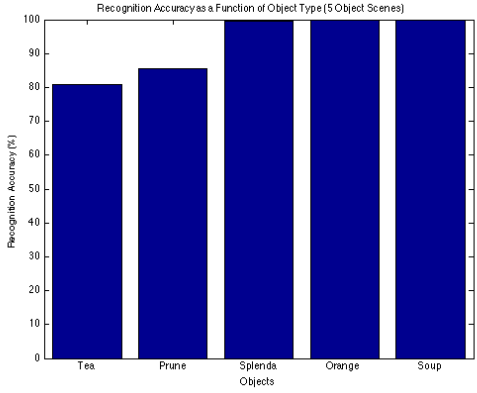

Fig. 2. Plot of true positive recognition rate for each of the five objects in 5-object scenes.

Scenes with 10 household objects exhibit an expected reduction in accuracy due to the presence of more objects in the same field of view, resulting in lower resolution for each object. Accuracy for 5 of the objects ranges from 63% to 96%. However, true positive recognition rates of the other 5 objects consistently remain less than 50%, with three of these below 10% chance level. Thus, attention biasing is helpful for some objects (Orange, Prune, Soup, Splenda, Cereal). These also are the largest of the 10 representative household objects chosen; hence, they fill most of the area in the recognized subwindows, making them well-matched to the subwindow size chosen (180 by 180 pixels). By contrast, the 5 objects for which attention-biased object recognition exhibits low accuracy occupy only a fraction of the chosen subwindow, suggesting that conducting these analyses with tuned subwindow sizes may improve accuracy for certain objects, as less surrounding clutter will reduce confusion for the feature-based recognition algorithm. Therefore, future extensions will harness depth information via input from a PrimeSense depth sensor to obtain more accurate object segmentation, in order to filter objects based on size and shape, enabling more robust, accurate object recognition.

3.3 Object Localization and Tracking System (OLTS)

In the context of this system, tracking algorithms serve as a means to detect objects in real time. Real time conditions come in to play because one of the main purposes of the system is to give dynamic feedback to human users. Thus,

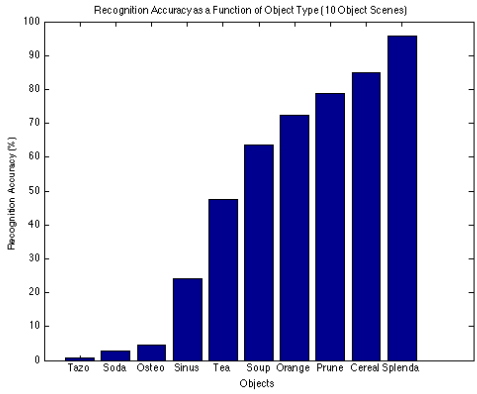

Fig. 3. Plot of true positive recognition rate for each of the ten objects in 10-object scenes. Seven of the ten objects exhibit classification accuracy above chance (ten percent).

it was necessary to choose a state of the art tracker. There are several trackers which track objects via assuming search areas within the frame [29, 30, 31], or which use state prediction via particle filters to determine object trajectory.

The aforementioned trackers yield desirable results; however, their flaws are exposed in videos exhibiting abrupt motion, frame-cuts, or objects leaving the field of view [32]. These are all constraints necessary for the OLTS. While employing the OLTS, the user wears a head mounted camera with a 100 degree field of view (FOV); thus, abrupt motion or objects leaving the FOV may occur. These considerations led us to utilizing the Context Tracker [32].

The system utilizes the Context Tracker and an auditory feedback module, the Sound Map. The system only requires one video frame with a bounding box enclosing the object to be tracked. Once this bounding box has been provided to the tracker, the tracker continually detects the object in the following frames, and the position of the object is updated within the control loop of the program. This position is then passed to the auditory feedback algorithm. The feedback algorithm then provides speech-synthesized commands to the user based on the position of the aforementioned object. The specifics of the vision algorithm and feedback module will be explained below.

Context Tracker The Context Tracker, which makes use of the P-N tracker [32], was the tracker of choice for the OLTS, as it is a basic target tracker. In addition to P-N learning, the Context tracker utilizes contextual information to robustly track objects. This contextual information is categorized into two entities: supporters and distracters. Supporters are features that consistently occur around the object; distracters are regions that have similar appearance to the actual target/object [4]. The addition of supporters and distracters allows the context tracker to deal with frame cuts, similar objects, and abrupt motion.

Fig. 4. OLTS hardware. A camera mounted to glasses sends visual input to the computer. The computer houses the vision and feedback algorithms. The bone conduction headphones relay information to the subject.

Auditory Feedback The feedback algorithm, the Sound Map, yields speech-synthesized feedback to the user based on the position of the desired object. It does so by discretizing the camera field of view into 9 regions (Figure 5). Depending upon the region in which the object resides, the computer synthesizes spoken words back to the user.

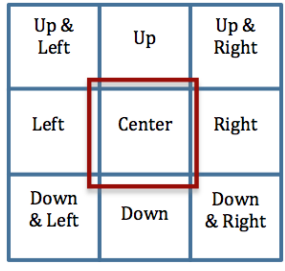

Fig. 5. Sound Map for auditory feedback mechanism. The grid represents a camera field of view (FOV). The position of the object in 3D space is mapped to the 2D FOV above. Once the object is mapped to this FOV, it falls into one of the 9 grid cells. Depending upon the grid to which the object belongs, the computer conveys the corresponding word. The size of this entire grid covers 640 x 480 pixels.

The user's goal is to listen to voice feedback, and turn his or her head based on the synthesized words. Once the object is within the "Center" region of the camera, the computer conveys the synthesized word "Center" to the user through bone conduction headphones. The user then reaches and grasps the desired object. A central angle of 23.4 degrees was chosen, as this was determined to allow for optimal grasp based on another study [33].

4 System Integration

The modular components of the system were integrated within the Neuromorphic Robotics Toolkit developed at the University of Southern California [34]. The attention-biasing, recognition, user interface, tracking, and feedback modules are shown below in Figure 6 with their corresponding input and output ports. These ensure correct information is passed between modules as needed to allow for their functionality, much like arguments can be passed from function to function in a computer program.

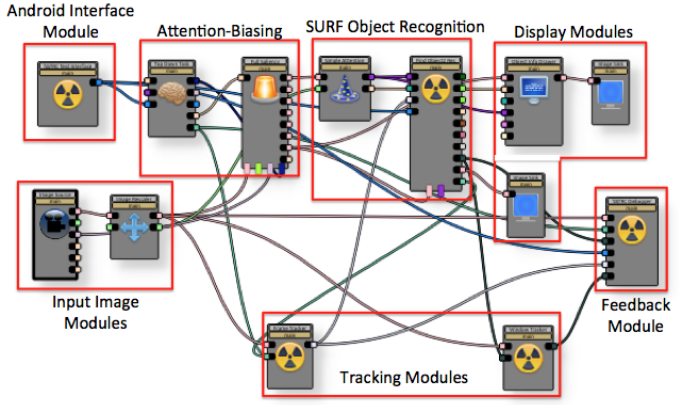

Fig. 6. System integration in Neuromorphic Robotics Toolkit; system components are labeled; links between modules represent inputs and outputs needed for functioning of each subpart based on information from other subparts.

5 Methods

The prototype system was taken to the Braille Institute in Los Angeles, California (USA) for testing with visually impaired subjects. Subjects were seated in front of one, two, or three household objects either positioned at the center, left, or right of the visual field. They were provided with an Android phone interface with the "Talk Accessibility Mode" turned on (upon one tap of a button on the screen, the function of the button is spoken out loud; two quick taps in succession are required to execute the command controlled by the button). The Android interface employs software developed in [27], which allows the user to "Find" a specific item and then shows a screen of ten possible objects (images of the Android interface screens are shown in Figure 7); we restricted search in this experiment to a box of Splenda or a carton of Orange juice. Subjects were instructed to double-tap the name of a single pre-specified object in front of them and then wait to hear cues as to where the object was located; when

the 'center' or 'stay' command was given, they were instructed to reach out with their hand in a straight line from their nose until they touched the object situated approximately arm's length (0.45m to 0.61m, 1.5ft to 2ft) away. It is noteworthy that currently the computer vision algorithms are implemented on a laptop, for development and testing purposes, that may be placed in a backpack worn by the user for mobility; work is ongoing toward designing and building dedicated hardware for the vision algorithms described that is even smaller and more portable.

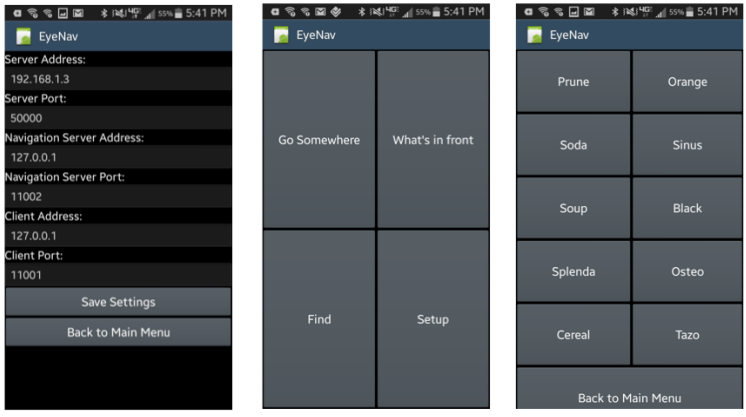

Fig. 7. Images of Android phone interface screens: (left) control for setting IP address and port number for communication with server module described in Section 3.1, (middle) option to 'Find' a specific object, (right) specific object query options for localization and grasp.

Fig. 8. (Left) Subject wearing the head mounted camera and bone conduction headphones. (Right) Example scene with objects and 9-region grid overlayed.

6 Results

Three visually-impaired subjects (due to Retinitis Pigmentosa, Cytomegalovirus Retinitis, and Optic Nerve Dysplasia, respectively) utilized the system and provided their evaluation using the System Usability Scale [35]. Reported scores were 82.5, 92.5, and 80.0, respectively. Time was recorded for each stage of the algorithm for two of these subjects (an optimal method of data collection was established after the first subject), including time to grasp the object and total time to use the system from the time the query was initiated. Average time is reported below in seconds for each stage (AB: Attention-Biasing). Out of the ten trials conducted for each subject, trials were excluded in cases when retraining was required mid-use due to a misunderstanding of how to use the system, and trials were excluded if the recognition algorithm failed so as not to confuse the subjects (this will be discussed in further detail in the next section).

Table 1. Time taken for Attention-Biasing (AB), Recognition, and Grasping (in seconds). Of the 10 trials conducted, 4 were excluded from subject RP and 2 from subject RT (2 of these (for RP), 1(for RT) due to subject misinterpretation of cues and 2 (for RP), 1(for RT) due to incorrect recognition result from algorithm).

Patient ID	AB	Recognition	Grasp	Total
RP (n = 6)	6.20	5.55	13.0	31.6
RT (n = 8)	7.36	7.43	12.8	41.8

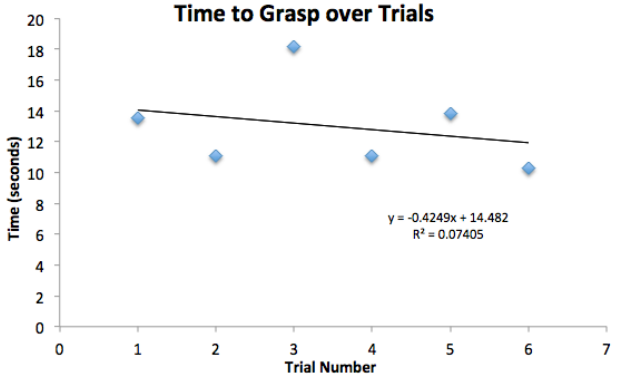

Fig. 9. Time taken to grasp object as a function of trial number for subject RP, p = 0.602.

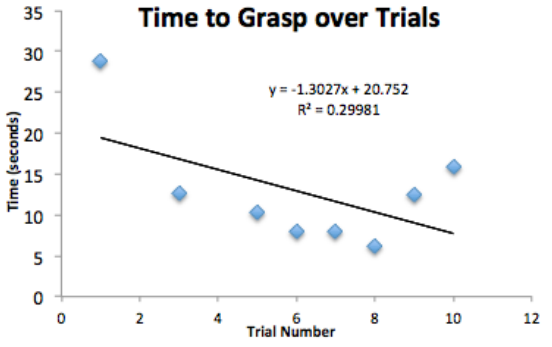

Fig. 10. Time taken to grasp object as a function of trial number for subject RT, p = 0.160.

7 Discussion

From results shown above, we observe that subjects were able to use the system independently for a majority of the conducted trials and successfully obtain the desired object. Out of the 10 trials conducted for each subject, in 2 out of the 10 trials for subject RP and 1 out of 10 for subject RT, the recognition algorithm did not correctly recognize the presence of the object; hence these trials were excluded. In the case of RP, for 2 trials, the subject extended a hand outward higher than nose level, so these were excluded; the subject was retrained before experiments were continued. In the case of RT, for 1 trial, the subject moved a hand laterally to one side instead of straight out in front for one trial, so the subject was retrained before continuing. This could be attributed to the fact that subject RT was blind from birth and thus proprioception of space and positions of body parts may not be as intuitive. The difference in time for achieving recognition (purely an algorithmic output) can be explained by the fact that the recognition process is sequential (each bounding box obtained from the attention-biasing step is compared to the database of trained objects; the final box selected as the recognition result is the one for which the matched object has the same label as that of the query requested, so depending on which of the three boxes is correct, the algorithm must process from 1 to 3 boxes). Furthermore, timestamps were taken by human hand with a stopwatch. In future experiments, these will be logged automatically by the computer being used to run the algorithms. For both subjects, no learning effect was observed ($p > 0.1$ in both cases) over trials with the system. This insignificant learning trend can be attributed to the extensive training completed with the subjects prior to running these experiments. Conducting experiments with more subjects and with more trials per subject will help to confirm this conclusion.

8 Conclusions

Our prototype system provides visually-impaired users the ability to query for a specific item of interest and receive explicit object recognition, localization, and feedback information to allow for ease of grasping of the desired object. To our knowledge, this is the first closed-loop system of its kind that provides explicit object localization and recognition as well as audio feedback for grasp without the need for the user to define a relevant region of the scene. The average time to grasp a desired object is between 12 to 13 seconds from recognition response (31 to 42 seconds from query initiation). These results along with the high system usability scale scores given by these three subjects indicate that this assistive computer vision tool is a promising aid for the visually-impaired, who make up nearly 5% of the global population. It is expected that the grasp response times can be improved to under 10 seconds with the incorporation of real-time hardware implementations of the algorithms described here, such as FPGA or GPU based implementations [36]. Furthermore, the computer vision algorithms described have the potential to serve not only as assistive algorithms for the blind but can also provide autonomous recognition and tracking in defense or exploratory (underwater/space) situations, that may be unfit or unsafe for human travel.

9 Acknowledgements

The authors would like to thank Carey Zhang and Sophie Marat for their essential algorithmic contributions to the development of the prototype system described in this manuscript.

References

[1] Visual Impairment and Blindness Fact Sheet, "World Health Organization," [online], 2012, http://www.who.int/mediacentre/factsheets/fs282/en/. [Accessed: 6-May-2013].

[2] A. C. Nau, "Gaps in Assistive Technology for the Blind: Understanding the Needs of the Disabled." Keynote Lecture, *IEEE ICME Workshop on Multimodal and Alternative Perception for Visually Impaired People (MAP4VIP)*, San Jose, CA, July 2013.

[3] R. Manduchi and J. Coughlan, "The Last Meter: Blind Visual Guidance to a Target," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[4] R. Manduchi and J. Coughlan, "(Computer) vision without sight," *Communications of the ACM*, vol. 55, no. 1, 2012.

[5] K. Thakoor, S. Marat, P. J. Nasiatka, B. P. McIntosh, F. E. Sahin, A. R. Tanguay, J. D. Weiland, L. Itti, "Attention-Biased Speeded-Up Robust Features (AB-SURF): A Neurally-Inspired Object Recognition Algorithm for A Wearable Aid for the Visually Impaired," in *IEEE ICME Workshop on Multimodal and Alternative Perception for Visually Impaired People (MAP4VIP)*, San Jose, CA, July 2013. Best Student Paper Award.

[6] M. Bjorkman, J-O. Eklundh. "Vision in the Real World: Finding, Attending, and Recognizing Objects." *International Journal of Imaging Systems and Technology.* vol. 16, pp. 189–208, 2007.

[7] B. Schauerte, M. Martinez, A. Constantinescu, "An Assistive Vision System for the Blind that Helps Find Lost Things," in *Proceedings of the 13th International Conference on Computers Helping People with Special Needs.* vol 2, pp. 566–572, 2012.

[8] J. P. Bigham, C. Jayant, A. Miller, B. White, T. Yeh, "VizWiz:: LocateIt-enabling blind people to locate objects in their environment," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.

[9] S. C. Nanayakkara, R. Shilkrot, P. Maes,"EyeRing: A Finger-worn Assistant," in *International ACM SIGCHI Conference on Human Factors in Computing*, Austin, TX, 2012.

[10] K. Matusiak, P. Skulimowski, and P. Strurnillo, "Object recognition in a mobile phone application for visually impaired users," in *The 6th International Conference on Human System Interaction (HSI)*, pp. 479–484, 2013.

[11] Looktel Recognizer, "Looktel," [online] 2009, http://www.looktel.com/recognizer (Accessed 23 February 2013).

[12] "OrCam - See for Yourself." [Online]. Available: http://www.orcam.com/. [Accessed: 01-May-2014].

[13] J. M. Wolfe. "Guided search 2.0: a revised model of visual search." *Psychonomic Bulletin and Review*, vol. 1, Issue 2, pp. 202–238, 1994

[14] A. M. Treisman and G. Gelade. "A Feature-Integration Theory of Attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[15] A. R. T. Gepperth, S. Rebhan, S. Hasler, J. Fritsch,"Biased Competition in Visual Processing Hierarchies: A Learning Approach Using Multiple Cues" *Cognitive Computation*, vol. 3, no. 1, pp. 146–166, Mar. 2011.

[16] T. Winlock, E. Christiansen, and S. Belongie, "Toward real-time grocery detection for the visually impaired," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 49–56, 2010.

[17] P. B. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

[18] E. Striem-Amit, M. Guendelman, and A. Amedi, "Visual Acuity of the Congenitally Blind Using Visual-to-Auditory Sensory Substitution," *PLoS ONE*, vol. 7, no. 3, Mar. 2012.

[19] C. Papageorgiou and T. Poggio. "A trainable system for object detection." *International Journal of Computer Vision*, vol. 38, Issue 1, pp. 15–33, 2000

[20] B. Moghaddam and A. Pentland. "Probabilistic visual learning for object representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, Issue 7, pp. 696–710, 1997.

[21] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guerin-Dugue, "Modeling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos." *International Journal of Computer Vision*, vol. 82, Issue 3, pp. 231–243, 2009.

[22] U. Rutishauser, D. Walther, C. Koch, and P. Perona. "Is bottom-up attention useful for object recognition?" *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[23] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, Jan. 2005.

[24] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[25] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[26] H. Bay, T. Tuytelaars, L. Van Gool, "SURF: Speeded Up Robust Features," in *European Conference on Computer Vision*, Graz, Austria, 2006.

[27] A. Adebiyi, C. Zhang, K. Thakoor, J. D. Weiland, "Feedback measures for a wearable visual aid designed for the visually impaired." *Association for Research in Vision and Ophthalmology Annual Meeting*. Seattle, Washington, May 5–9, 2013.

[28] M. Aly, P. Welinder, M. Munich, P. Perona, "Scaling object recognition: Benchmark of current state of the art techniques," in *IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, 2009.

[29] A. Adam, E. Rivlin, and I. Shimshoni. "Robust fragments- based tracking using the integral histogram." in *International Conference on Computer Vision and Pattern Recognition*, 2006.

[30] B. Babenko, M.-H. Yang, and S. Belongie. "Visual tracking with online multiple instance learning." in *International Conference on Computer Vision and Pattern Recognition*, 2009.

[31] H. Grabner, C. Leistner, and H. Bishof. "Semi-supervised on-line boosting for robust tracking." in *European Conference on Computer Vision*, 2008.

[32] T. Dinh, N. Vo, G. Medioni. "Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments." In *International Conference on Computer Vision and Pattern Recognition*, 2011.

[33] N. Mante, G. Medioni, A. Tanguay, J. Weiland. "An Auditory Feedback Study on the Object Localization and Tracking System." in *Biomedical Engineers Society Annual Meeting (BMES Annual Meeting)*, 2014.

[34] "iLab Neuromorphic Robotics Toolkit: Get NRT." [Online]. Available: http://nrtkit.org/documentation/g_GetNRT.html. [Accessed: 29-Jun-2014].

[35] "Measuring Usability with the System Usability Scale (SUS): Measuring Usability." [Online]. Available: http://www.measuringusability.com/sus.php. [Accessed: 29-Jun-2014].

[36] S. Kestur, M. S. Park, J. Sabarad, D. Dantara, V. Narayanan, "Emulating mammalian vision on reconfigurable hardware," in *IEEE 20th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pp. 141–148, 2012.