ATTENTION BIASED SPEEDED UP ROBUST FEATURES (AB-SURF): A NEURALLY-INSPIRED OBJECT RECOGNITION ALGORITHM FOR A WEARABLE AID FOR THE VISUALLY-IMPAIRED

Kaveri A. Thakoor, Sophie Marat, Patrick J. Nasiatka, Ben P. McIntosh,
Furkan E. Sahin, Armand R. Tanguay, Jr., James D. Weiland, and Laurent Itti
University of Southern California
thakoor@usc.edu, sophie.marat.ilab@gmail.com, nasiatka@usc.edu, bmcintos@usc.edu,
fsahin@usc.edu, atanguay@usc.edu, jweiland@med.usc.edu, itti@usc.edu

ABSTRACT

Humans recognize objects effortlessly, in spite of changes in scale, position, and illumination. Emulating human recognition in machines remains a challenge. This paper describes computer vision algorithms aimed at helping visually-impaired people locate and recognize objects. Our neurally-inspired computer vision algorithm, called Attention Biased Speeded Up Robust Features (AB-SURF), harnesses features that characterize human visual attention to make the recognition task more tractable. An attention biasing algorithm selects the most task-driven salient regions in an image. Next, the SURF object recognition algorithm is applied on this narrowed subsection of the original image. Testing on images containing 5 different objects exhibits accuracies ranging from 80% to 100%. Furthermore, testing on images containing 10 objects yields accuracies between 63% and 96% for the 5 objects that occupy the largest area within the image subwindows chosen by attention biasing. A five-fold speed-up is attained using AB-SURF as compared to the time estimated for sliding window recognition on the same images.

Index Terms *— Object recognition, visual attention, neurally-inspired computer vision, visual aids for the blind*

1. INTRODUCTION

1.1. Background

Object recognition and localization are key perceptual tasks with which blind people often struggle [1]. Over the past decade, numerous Electronic Travel Aids have been developed that aid in scene interpretation with sound [2], navigation and object localization [3], and even object or text recognition [4–6]. Several such devices rely on stereo and depth information to create 3D models of the environment [7]. Others function by means of text-to-speech-based conversion technologies. However, these prior technologies do not expressly address search for and recognition of specific items desired by the user [8]. The novelty of the algorithm sequence described herein lies principally in its inspiration from the neural characteristics of human visual attention, including orientation, color, and intensity, to reduce the space over which the SURF recognition algorithm must then compute to locate a queried object. Once validated, these algorithms will be integrated with mobile computing and auditory/tactile feedback (as described in [9]) into a Wearable Visual Aid to provide visually-impaired users with explicit object recognition in real time.

1.2. Wearable Visual Aid System Overview

The Wearable Visual Aid system is composed of six key components, as shown in Fig. 1 below.

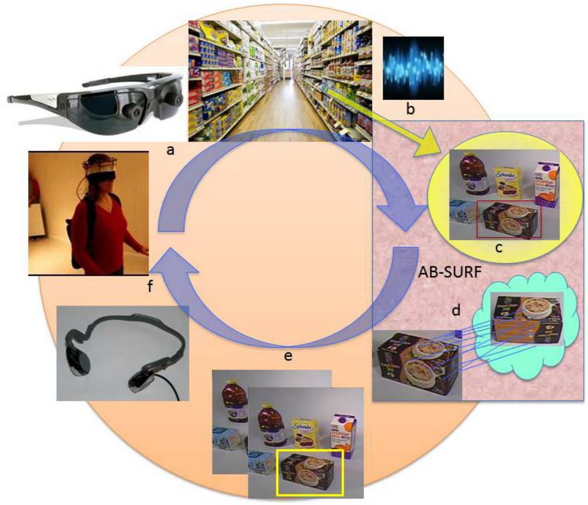

Fig. 1. System Overview: (a) Head-mounted camera, integrated within glasses, to capture the external scene; (b) Voice command module for user query; (c) Attention biasing to narrow the region of interest (shown in red box); (d) Queried object recognized by comparison with the best match from the database (image shown in cloud), [(c) and (d) together comprise AB-SURF]; (e) Confirmed object location marked in yellow box, subsequently tracked over time in multiple frames; and (f) Audio cues provided back to the user via bone conduction headphones.

The system operation is as follows. First, images of the external scene are captured by a head-mounted, wide-field-of-view camera. A voice command module allows the user to speak a request for a specific desired item. Next, neurally-inspired computer vision algorithms locate and recognize the desired object. An attention biasing algorithm narrows down to the region of the current scene that is most likely to contain the desired item. An object recognition algorithm serves as the gate, confirming whether or not the object in the narrowed region is what the user requested. After the object has been located, position and dimension information are sent to an object tracker. Finally, an audio feedback module provides the user with cues to help maintain the object in the user's field of view until he or she successfully grasps it.

2. RELATED WORK

Recognition aids for the blind are pervasive and in demand; they are now available as Smartphone applications, with IQ Engine's oMoby [10] (which makes use of real-time crowd-sourcing when an object is encountered that is not present in the trained database) and LookTel's Recognizer [5]. The EyeRing [4] can recognize dollar bills, and one proposed system is able to help visually impaired people in choosing the perfect party attire [6]. At the level of landmark localization, the authors of [11] detect restroom door signs and elevator buttons using a combination of saliency followed by the sliding window technique and Bipartite Graph Matching. However, the sliding window technique is computationally intensive. By contrast, a technique that combines 3D laser and reflection imagery data along with visual attention to recognize objects has been designed and evaluated for mobile robotics applications [12]. Although the utilization of laser and reflectance imagery together adds robustness to the results obtained, it also adds significant cost and may not be compatible with a compact, wearable system.

Our approach presents several key advantages. First, it involves less extensive computation, as a consequence of employing biased attention, than a sliding window object recognition approach. With the aid of neurally-inspired attention biasing, we can isolate regions of interest within a scene of multiple objects and perform real-time recognition on these biased subwindows, achieving a speed-up of greater than five-fold as compared with the brute force sliding window approach. Second, we achieve functional simplicity compared to laser/reflectance approaches, as we utilize a single (non-stereo, RGB) camera for input. By using a glasses-mounted wide-field-of-view camera, we enable scene scanning in a natural way for the visually-impaired user. Third, we recognize objects within a complete scene containing distractors, as compared to several Smartphone recognizer applications (that require physically isolating the object to be detected in the field of view of the camera). We evaluate the algorithm on sets of 5 and 10 household items (exemplars for the potential objects a blind user may encounter while searching for an item in his or her pantry or at the supermarket) as a simplified testbed to judge performance of attention-biased object recognition before extending to field images in an actual user's home or supermarket.

This paper is organized as follows. Section 3 describes the use of scenes with 5 or 10 objects on a plain background to test the novel attention biasing algorithm, and Section 4 focuses on the determination of which object recognition algorithm to use for the full system by evaluation on images of isolated objects one at a time. Section 5 then details the evaluation of a fully integrated AB-SURF system that chooses the top 3 candidate objects using attention biasing, recognizes them, and selects the best recognition result as the final output.

3. ATTENTION BIASING: NARROWING THE SEARCH

3.1. The Role of Attention Biasing

A set of features based on hue, orientation, and intensity are extracted from an input image to select regions of interest given a user's query. The extracted features are then weighted based on the object of interest (the user's search query), allowing biasing based on task (*top-down*), rather than strictly based on bottom-up salience (which weights all features equally). Example attention-biased saliency maps (saliency encoded as brightness plotted as a function of image location) are shown in Fig. 2. This first stage of attention biasing pinpoints the region of the image most likely to contain a desired object, reducing the space over which object recognition must compute. Attention biasing enhances the computational efficiency by subduing distracting objects, even though they may be more inherently salient (as can be seen by the distinctly different highlighted regions in the biased as compared with the unbiased saliency maps presented in Fig. 2).

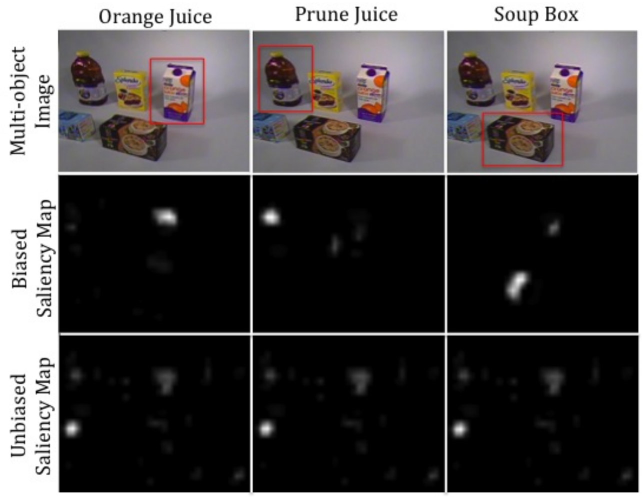

Fig. 2. Example of top-down attention biasing. Red boxes surround desired objects, to which attention has been focused in the biased saliency map, as compared to the unbiased saliency map, which highlights all salient objects equally.

3.2. Computation of Attention-Biased Saliency Map

The attention biasing algorithm is based on the theory developed in [13]. The attention biasing algorithm learns the statistics of the desired object and its distractors at the level of different feature maps that are combined to form the final saliency map. During the test phase, the relevant features for a particular object are enhanced, increasing the saliency of the region containing the desired object. The biasing relies strongly on hue and orientation, for which 24 and 8 features were extracted, respectively. Unlike Refs. [12–15], intensity and saturation were not used, as they were not relevant in our current testing to distinguish one item from another. They may be more useful in distinguishing various objects from a cluttered background rather than from other objects.

3.3. Attention Biasing Results

The different parameters involved in the building of the saliency map (normalization and number of channels used, for example) were investigated by filming a video presenting 10 household items (200 training images and 380 testing images, with an example shown in Fig. 4) and a video presenting 5 household items (100 training images and 660 testing images, with an example shown in Fig. 7) and then extracting individual frames from each video. We evaluate both the best location chosen by the algorithm (1 location) and the three best locations chosen by the algorithm (3 locations). For the 3-location case, the answer is considered as a true positive answer if at least one of the three locations falls on the desired object. For the 5-object case, using only one location is already very efficient; all of the objects are found more than 85% of the time. Using 3 locations guarantees correct fixation at greater than 96% for all objects. For the 10-object case, using only one location is already sufficient for correctly locating some objects (orange juice carton, prune juice bottle, box of Splenda, and box of tea bags) with results in the range of 88% to 99% (Table 2). The results increase even further (98% to 100%) when three locations are considered. On the other hand, attention biasing on several other objects is not as predictive with one location (8% to 45%), while consideration of three potential locations improves this rate (44% to 94%). Clearly, increasing the number of potential locations enhances the chances of success, showing marked improvements in locating those difficult objects that are not easily biased by the algorithm with just one location. A sample confusion matrix for the 5-object, 3-location case is presented in Fig. 3. The percentages of correct fixations on desired objects for 5 and 10-object cases are shown in Tables 1 and 2, respectively.

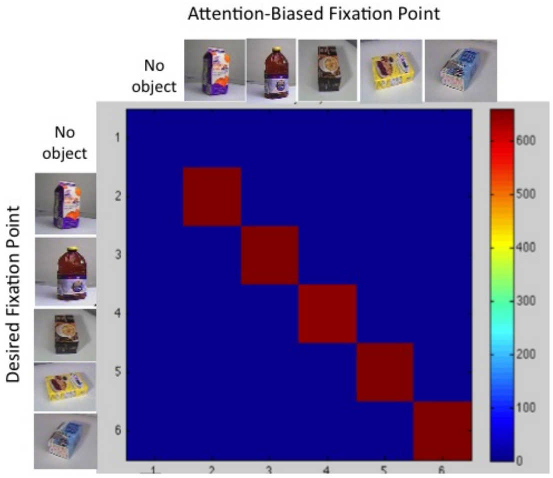

Fig. 3. Evaluation of attention biasing on 5 household objects for 3 fixation locations. The maximum number of correct fixations possible was 660, indicated by the darkest shade of red.

Table 1. Percentages of correct fixations for attention biasing on 5-object images.

	Orange	Prune	Soup	Splenda	Tea
1 location	99%	88%	85%	95%	100%
3 locations	99%	99%	96%	99%	100%

Table 2. Percentages of correct fixations for attention biasing on 10-object images.

	Cereal	Orange	Prune	Sinus	Splenda
1 location	45%	98%	95%	34%	88%
3 locations	94%	99%	100%	91%	98%
	Osteo	Soda	Soup	Tazo	Tea
1 location	8%	21%	26%	31%	99%
3 locations	44%	65%	65%	47%	100%

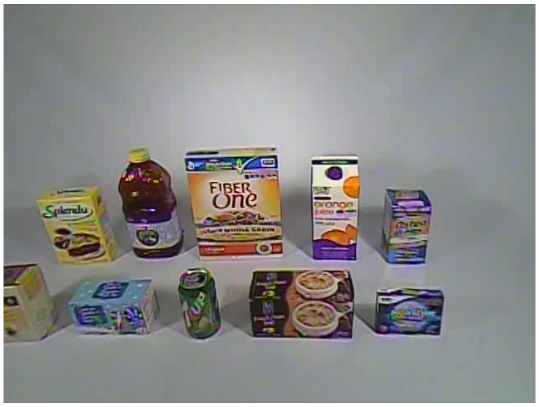

Fig. 4. Example image of 10 objects together.

4. OBJECT RECOGNITION

4.1. Comparing SURF and SIFT Algorithms

Two leading object recognition algorithms were evaluated for their potential incorporation into the wearable visual aid. David Lowe's SIFT (Scale Invariant Feature Transform) [16] and Bay, *et al.*'s SURF (Speeded-Up Robust Features) [17] were implemented within a custom-developed, modular C++ toolkit, both employing the functionality of Willow Garage's OpenCV Library. These algorithms were evaluated on images of single objects in isolation without any attention biasing to determine which algorithm performs better for recognition of household objects. SIFT recognizes objects in spite of variations in scale, rotation, and translation, and is also partially invariant to changes in illumination and affine projection. SURF is a faster, and in some cases more accurate, alternative to SIFT. Both achieve their tasks by first extracting keypoints or "interest points" by convolving with Gaussian filters (or an approximation of Gaussian filters, *e.g.* box filters, in the case of SURF), followed by creating descriptors of these keypoints based on their orientation and position information, and finally by matching query and database images that are closest in multidimensional Euclidean distance.

4.2. Methods for Comparing SURF and SIFT

Comparison testing for these two algorithms was conducted separately on 10 household items in isolation. In one set of experiments used to generate the Receiver Operating Characteristic (ROC) curve shown in Fig. 5, training was performed on 500 instances (images of one object at varying angles) of one object at a time, followed by performance testing on 600 images, comprising 60 different instances of each of the 10 objects (including different views of the object used for training and different views of nine untrained objects). A centralized rotational platform with neutral backgrounds in conjunction with a suite of custom illumination controls and multiple KPC-E23NUB wide-dynamic-range NTSC cameras provided for uniform image acquisition of the test objects. Each individual test object was recorded from multiple camera positions and with single and multiple light sources enabled/disabled during video recording, all while the objects were rotated on a precision rotary table. Captured image sequences were recorded at 30 frames per second using a 4-channel Sensoray NTSC converter box interfaced to a PC via a USB 2.0 interface and de-interlaced before saving in uncompressed AVI video format. In another set of experiments designed to judge which objects are recognized best by these two algorithms by generating the confusion matrix shown in Fig. 6, separate testing was performed, this time with a training set of 600 images (60 different instances of each of the 10 objects), followed by a testing set of 5,000 images (500 different instances of each of the 10 objects). Algorithm speed was also monitored by recording the time taken for database training. Results of both of these experiments are described in the following two sections.

4.3. SURF Algorithm Performance

SURF outperforms SIFT in recognition accuracy for all 10 objects tested in this phase, as can be seen from the larger area under the curve in the exemplifying ROC curve for one sample object (Fig. 5).

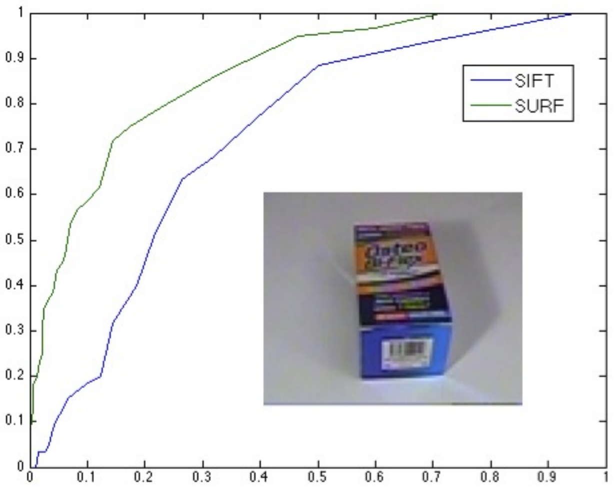

Fig. 5. Example Receiver Operating Characteristic (ROC) curve for a sample grocery store item; SURF outperforms SIFT.

The difference in training time was another significant distinguishing factor between the two algorithms. Average database training time was over an order of magnitude faster for SURF as compared to SIFT (43 seconds for SURF and 514 seconds for SIFT). The increased recognition performance and speed of training for SURF may be attributed to the lower dimensionality of the SURF descriptor as well as the use of image processing via integral images; these two advancements afford SURF with faster as well as more robust household object classification for this application. For these reasons, SURF was chosen as the algorithm to be incorporated into the fully integrated system.

4.4. SURF Object Recognition Accuracy: "Not All Objects are Created Equal"

Testing SURF object recognition on 10 household objects indicates that the best-recognized objects include large boxes (Cereal, Soup), cartons (Orange), and flat-surfaced bottles (Prune) with brand names or logos clearly visible, allowing the algorithm to more easily retrieve highly discriminating features. Objects that were recognized poorly were smallest in size and hence made up a smaller portion of the field of view, thereby appearing at lower resolution in the final images than their larger counterparts and allowing for fewer keypoints to be extracted for SURF matching. The most poorly recognized objects exhibit significant glare (*e.g.*, Soda) in addition to being small. Examples of these ten objects and a confusion matrix displaying classification ability are shown in Fig. 6.

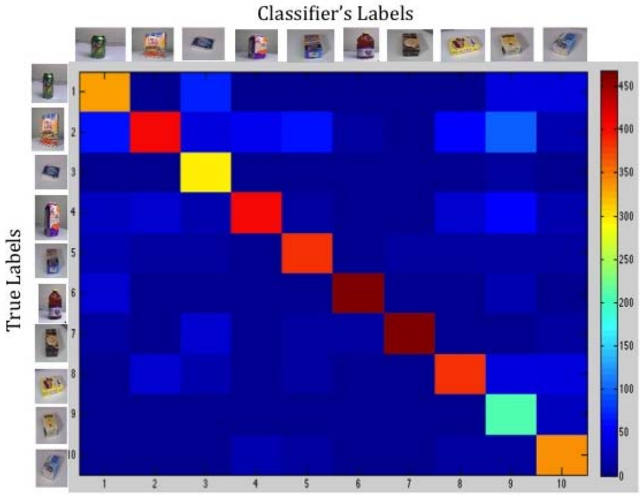

Fig. 6. Confusion matrix for SURF object classification; 500 objects of each kind were actually present in the testing phase.

Nevertheless, the accuracy (as defined in [18]),

$$A = (TP + TN)/(TP + TN + FP + FN), \qquad (1)$$

(in which A represents accuracy, TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives) of SURF object recognition is greater than 90% for all 10 objects, making it an effective algorithm to use for this system. Accuracy rates are shown for all 10 objects in Table 3.

Table 3. Isolated SURF object recognition performance (accuracy) on 10 objects.

Orange	Prune	Soup	Splenda	Tea
94.1%	97.9%	97.5%	94.6%	95.3%
Cereal	Osteo	Sinus	Soda	Tazo
90.1%	95.7%	95.5%	93.5%	93.3%

5. RESULTS FOR ATTENTION-BIASED OBJECT RECOGNITION

5.1. System Integration

The modules that implement attention biasing as well as object recognition have been integrated within a custom-developed modular toolkit, allowing for testing of the pipeline for attention-biased recognition. Three potential regions of high saliency, biased by the user's command, are provided to the object recognition algorithm (as shown in Fig. 7), out of which only the best recognition match (in agreement with the user's command) is returned for further processing and object tracking, as shown in the sample pipeline (Fig. 8). If the object is not recognized among the three narrowed rectangles, a message is displayed that the object has not been found, and program flow returns to the attention biasing module, waiting for another voice command or a new input image for analysis.

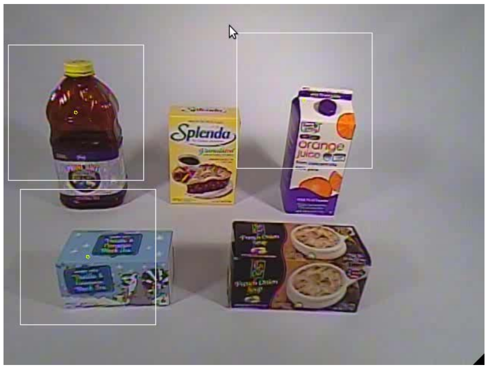

Fig. 7. Example 5-object image with the 3 most salient attention-biased regions (biased based on the user's query of orange juice) shown in white rectangles.

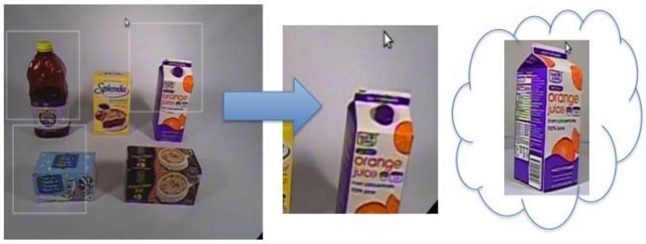

Fig. 8. SURF object recognition applied to 3 rectangles chosen by attention biasing; the best rectangle (right of blue arrow) was chosen based on the closest match to the database instance (depicted in the cloud) of the user query for orange juice.

5.2. AB-SURF: Algorithm Initial Results

Performance of attention-biased object recognition was tested on 655 (5-object) and 382 (10-object) images (sequential video frames taken from head-mounted camera view) after training the object recognition algorithm on 300 (5-object) and 600 (10-object) images (60 instances of each of the objects shown in Fig. 4 and 7), respectively. During the algorithm run, cropped boxes were extracted based on center points provided by the attention biasing algorithm. Recognition on these chosen subimages resulted in a determination of which box contained the desired object. As before, a wide-dynamic-range NTSC camera was used for image capture. The rectangle crop dimensions chosen for these experiments were 180 pixels by 180 pixels in 640-pixel by 480-pixel full-size images, but these crop dimensions could be scaled as desired for optimal recognition. Evaluation of the performance of the entire sequence (attention biasing followed by SURF recognition on cropped rectangles surrounding attention-biased points and resulting successful confirmation of the presence of the user's requested item) gives the true positive rates (*TPR*s) shown in Tables 4 and 5, as defined by

$$TPR = TP/(\text{total \# of images}), \qquad (2)$$

in which *TP* represents True Positives, and the total numbers of images were 655 and 382 for 5-object and 10-object images, respectively.

Table 4. Attention-biased object recognition true positive rates for 5-object images.

Orange	Prune	Soup	Splenda	Tea
99.8%	85.6%	100%	99.6%	80.9%

Table 5. Attention-biased object recognition true positive rates for 10-object images.

Orange	Prune	Soup	Splenda	Tea
72.3%	78.8%	63.6%	95.8%	47.6%
Cereal	Osteo	Sinus	Soda	Tazo
85.0%	4.45%	24.1%	2.89%	0.787%

All five objects present in the 5-object images exhibit true positive recognition rates of greater than 80% (compared to recognition of greater than 90% accuracy when recognizing these objects in isolation), with three of the five objects having recognition rates of greater than 99%. 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. In the 10-object images, recognition accuracy for 5 of the objects ranges from 63% to 96%. However, recognition of the other 5 objects consistently remains in the range of 0.79% to 47%. Overall, the 5-object case yields excellent results. The 10-object case exhibits 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. Thus, attention biasing is helpful for some objects (Orange, Prune, Soup, Splenda, Cereal). These also are the largest objects of the 10 representative household objects chosen, making them well-matched to the subwindow size chosen (180 by 180 pixels); they fill most of the area in these subwindows. 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 experiments with varying subwindow sizes may improve accuracy for certain objects. These lower rates are also partially attributable to the inherent limitations of the SURF algorithm, which could be addressed by using alternative recognition algorithms, such as the Deformable Parts Model (DPM) or Hierarchical Models of Object Recognition in Cortex (HMAX).

6. CONCLUSIONS AND FUTURE WORK

By pinpointing regions of an image that are most likely to contain a desired object using human-visual-attention-inspired features, the object recognition problem is made tractable. If we evaluated all possible recognition subwindows with 50% overlap, we would need to process at least 40 recognitions per frame; by using attention, we gain the advantage of processing only 3. The compute time for attention is about the same as 4 recognitions. Hence, the speedup is at least 40/(3+4), or 5.71 times faster, since the number of recognitions necessary will increase as subwindow overlap is increased to optimize performance. The AB-SURF algorithm has achieved recognition accuracy rates of 80% to 100% on 5-object images and accuracy rates of 63% to 96% on 5 objects (for which the chosen subwindow size for this experiment was optimal) in 10-object images, where the images consisted of several household objects, including representative boxes, cartons, and bottles that a blind user may encounter in his/her pantry or at a supermarket. However, some objects are not recognized well; this is a limitation of SURF that could be addressed by using alternative recognition algorithms (*e.g.*, HMAX, DPM, or Oriented Binary Robust Independent Elementary Features, ORB). Also, as the number of objects in the scene increases, attention biasing no longer improves the recognition rate; this could be addressed by adding more tuned feature extraction channels. Future work will involve optimization of both camera image sensor resolution and camera-object distance to obtain attention-biased recognition with higher accuracy. Experimenting with tuned window sizes is also expected to improve the recognition accuracy for certain objects. Testing will also be carried out on more realistic field images of objects in cabinets or on grocery store shelves. The algorithm may also be extended by incorporation of neural inspiration from computation of the "gist" of a scene. Based on these results, the AB-SURF neurally-inspired recognition algorithm is ready to be harnessed to provide information to tracking and audio feedback modules to direct users to a desired object until it is successfully grasped, paving the way for its integration into a deployable prototype Wearable Visual Aid system.

7. ACKNOWLEDGEMENTS

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. The authors would like to thank Dr. Gerard Medioni and Dr. Ram Nevatia for their insightful comments on this manuscript.

8. REFERENCES

[1] F. Dramas, *et al.*, "Artificial Vision for the Blind: A Bio-Inspired Algorithm for Objects and Obstacles Detection." *International Journal of Imaging and Graphics,* vol. 10, no. 4, pp. 531-544, 2010.

[2] E. Striem-Amit, *et al.*, "Visual Acuity of the Congenitally Blind Using Visual-to-Auditory Sensory Substitution." *PLoS ONE*, vol. 7, no. 3, 2012.

[3] About the UltraCane., "UltraCane: Putting the World at your Fingertips," [online] 2011, http://www.ultracane.com/about_the_ultracane (Accessed 3 February 2013).

[4] S. C. Nanayakkara, *et al.*, "EyeRing: A Finger-worn Assistant," in *International ACM SIGCHI Conference on Human Factors in Computing*, Austin, TX, 2012.

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

[6] F. Jabeen, *et al.*, "Intelligent Assistant to Help Blind People for Selecting Wearable Items," *Journal of American Science*, vol. 7, no. 8, 2011.

[7] A. Hub, *et al.*, "Interactive Tracking of Movable Objects for the Blind on the Basis of Environment Models and Perception-Oriented Object Recognition Methods" in *International ACM SIGACCESS Conference on Computers and Accessibility*, Portland, OR, 2006.

[8] D. Dakopoulos and N. G. Bourbakis, "Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 40, no. 1, pp. 25-35, 2010.

[9] A. Adebiyi, *et al.*, "Evaluation of Feedback Mechanisms for Wearable Visual Aids," in *Proc. IEEE ICME, Workshop on Multimodal and Alternative Perception for Visually Impaired People*, San Jose, CA, 2013.

[10] Technology, "Omoby," [online] 2006, http://www.iqengines.com/technology/ (Accessed 26 February 2013).

[11] W. Shuihua and Y. Tan, "Indoor Signage Detection Based on Saliency Map and Bipartite Graph Matching," in *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, Atlanta, GA, 2011.

[12] S. Frintrop, *et al.*, "Saliency-based Object Recognition in 3D Data," in *IEEE International Conference on Intelligent Robots and Systems*, Sendai, Japan, 2004.

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

[14] V. Navalpakkam and L. Itti, "An Integrated Model of Top-Down and Bottom-Up Attention for Optimizing Detection Speed," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009.

[15] L, Itti, C. Koch, E. Neibur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[16] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91-100, 2004.

[17] H. Bay, *et al.*, "SURF: Speeded Up Robust Features," in *European Conference on Computer Vision*, Graz, Austria, 2006.

[18] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers," HP Laboratories, Palo Alto, CA, 2004.