FROM RGB-D TO LOW-RESOLUTION TACTILE: SMART SAMPLING AND EARLY TESTING

Hao Tang[1,2], Margaret Vincent[3], Tony Ro[3] and Zhigang Zhu[2]

[1] Department of Computer Information System, CUNY Borough of Manhattan Community College
[2] Department of Computer Science, CUNY City College and Graduate Center
[3] Department of Psychology and CUNY Program in Cognitive Neuroscience
htang@bmcc.cuny.edu, margaretarlene@gmail.com, tro@ccny.cuny.edu, zhu@cs.ccny.cuny.edu

ABSTRACT

Because of low resolution and simple scene transducing approaches (e.g., uniform sampling) used in current visual prosthetic devices, only very limited restoration or substitution of vision can be provided to the visually impaired. Two interesting and challenging research problems arise. The first one is how to transduce the most important information to end users with very limited resolution; and the second one is how effectively visual prosthetic devices work with blind people. In this paper, various smart sampling and enhancement methods based on color and depth (RGB-D) segmentation are described that can be used to automatically select, sample and transduce the most useful scene information to end users with these visual substitution devices. The RGB-D segmentation is first generated by a patch-based stereovision algorithm, given a pair of color images captured by a stereo camera head. Then, the RGB-D segmentation is transduced into various alternative perception choices. Furthermore, some early studies on a visual prosthetic device using a tongue stimulator are described, and preliminary experiments are discussed that test the proposed methods.

Index Terms— image sampling, 3D reconstruction, assistive technology, visual prosthesis

1. INTRODUCTION

Based on the World Health Organization 2012 Report, there are 285 million visually impaired people in the world; 39 million are blind and 246 million have low vision. Using portable or wearable systems to assist the visually impaired for navigation has recently been the focus of much attention.

Various visual prostheses have been developed in the past decade. The most prevalent visual prosthesis, the retinal prosthesis, is an experimental visual device aimed at restoring vision functions of the visually impaired, and it has been studied and developed in many research groups and applied into some clinic trials [2, 16]. These retinal prostheses can provide low resolution images to a blind user that electrically stimulates his/her retinal cells. An example of such a device is Argus II from Second Sight [16], which has already received marketing approval. A retinal prosthesis is used to partially restore vision for blind and visually impaired people, especially for those who have lost their vision due to retinitis pigmentosa or macular degeneration. Currently, the state-of-the-art retinal prosthesis has very limited resolution (60 – 100 channels, in the form of a 6×10 or 10x10 array) [16].

The BrainPort V100 [1], developed by Wicab, Inc., is a non-surgical assistive device designed to assist blind and visually impaired individuals with object identification and navigation. The device uses a small video camera mounted on a pair of sunglasses to capture visual information, which is then converted into patterns of electrical stimulation on the tongue via a 20×20 grid electrode array. Users may be able to learn to perceive the shape, size, and location of objects in their environment through these patterns of tongue stimulation, hopefully allowing them to better understand their surroundings.

Both methods face a very serious problem: low resolution. If we simply sub-sample an original image into a 20×20 or lower resolution array to drive the retinal or tongue stimulation, it would be hard to identify small objects that are close to the user. Another problem is when a scene is cluttered and includes a lot of information and objects, in which case it is difficult to represent the complex scene in a low-resolution display. Furthermore, it is challenging to convey and enhance the most useful and important information without a comprehensive analysis of the scene.

In this paper we propose a set of novel 3D scene transducing methods for low image output resolution: (1) smart sub-sampling using both color and depth segmentation; (2) background removal based on scene labeling; (3) motion parallax simulation using 3D information; (4) dynamic object re-illumination and highlighting; and (5) traversable path direction estimation. We are performing experiments that send the transducing results to the tongue stimulator of the BrainPort device. We hope the new sampling and enhancement approaches will increase the capability of alternative perception techniques for the visually impaired.

The paper is organized as follows. Section 2 discusses some closely related work. An overview of our algorithm is given in Section 3. In Section 4, we present a number of smart subsampling and enhancement methods; Section 5 provides some experimental results in smart sampling. Sec-

tion 6 describes some of our early studies on using BrainPort for shape recognition and navigation. Finally we conclude our work in Section 7.

2. RELATED WORK

Because of recent advances in computer vision research, with a large body of various algorithms and their fast implementations, computer vision techniques are playing an increasingly important role in the development of visual prostheses [3]. Computer vision algorithms can be used to help restore some specific visual abilities, such as light perception and object recognition. Image segmentation has been applied in a visual prosthesis to enhance object recognition [5], and face detection and tracking methods are used to assist with recognizing faces [4]. Another challenge that visually impaired people encounter is navigation, and many different vision technologies have been applied in the development of electronic travel aids (ETA) for the visually impaired. Coughlan et al. [7, 8] propose systems for helping the visually impaired find a path to a machine-readable sign using a cellphone camera. Using stereo cameras [9, 10], depth maps are produced to aid navigation. Staircases [11, 12 and 13] and zebra-crossings [14] are detected using stereo cameras to help blind users identify and climb stairs and cross streets.

The work most related to ours is the method proposed by McCarthy et al. [6], which is a vision algorithm for a retinal prosthesis to support visual navigation. With stereo vision techniques, the system classifies a scene into ground and non-ground surfaces and renders a depth image in a low resolution version (<1000 pixels). The ground area is highlighted so it is easier for perception. The experiments show the down-sampling result in 2-bit and 6-bit dynamic range, but it might miss small/thin objects, such as a pole, a horizontal bar, or a thin tree branch in front of the user, due to the use of uniform sampling.

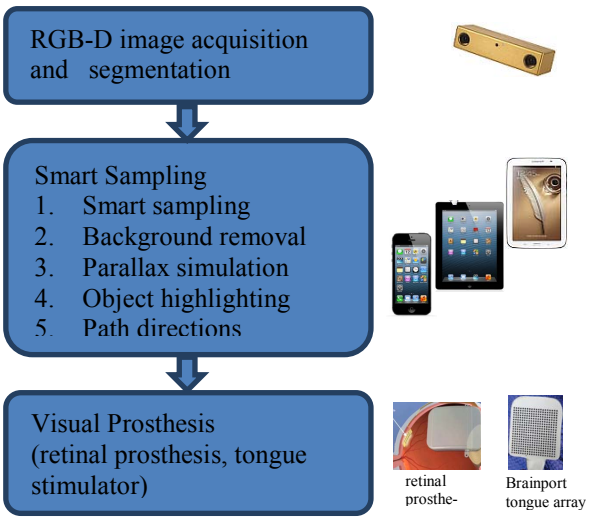

Fig. 1. The pipeline of a visual prosthesis system based on smart sampling.

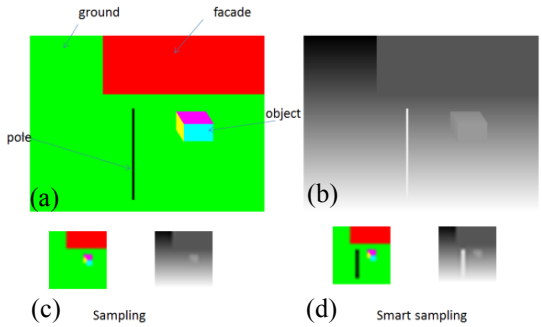

Fig. 2. Smart sampling based on 3D scene segmentation. (a) A color (RGB) image of the simulated scene with a number of objects (e.g., a pole is in a close range). (b) The depth map (D image) of the simulated scene. (c) Sampling results of the RGB and the D maps using a regular sampling method. Note the pole is missing after the regular sampling. (d) Sampling results of the RGB and D maps using our smart sampling method. The pole is still preserved after sampling.

3. OVERVIEW OF OUR APPROACH

The basic steps of a visual prosthetic system based on smart sampling are as follows (Fig. 1). First, images are captured with a stereo camera and the corresponding 3D model is recovered using a patch-based stereo vision method. Second, a number of smart sampling and object enhancement choices are applied, hence only the objects of interest (OIs) will be selected and sampled to generate low resolution images. Finally, the low resolution images are transduced to the low-resolution display for an end user who is visually impaired. The output of the system can be easily encoded into the input device of any kind of visual prosthesis that has low resolution, such as a retinal prosthetic or tongue stimulator.

Fig. 2 illustrates the basic idea of our smart sampling approach using a simulated scene. Fig. 2a is a simulated image with a green background (ground plane) and three objects: a building façade, a cubic object, and a thin, vertical pole. Using the patch-based stereo approach [15] based on color segmentation, the 3D information of all regions is obtained (Fig. 2b), and thus we have segmentation results of the scene with both color (RGB) and depth (D) information. Fig. 2c shows the results of a regular uniform sampling of the image into 20×20 pixels: the thin long pole disappeared. However, using our smart sampling method, the thin pole is preserved in the final 20×20 subsampled image.

4. SMART SAMPLING

The smart sampling approach consists of two steps. First, a color-patch-based stereovision method [15] is applied to a pair of stereo images captured by a stereo camera head. The outcome of the patch-based method is not just an array of individual 3D points that are usually produced by a typical stereovision system. Instead, it is a geometric representation of plane parameters, with geometric relations among neighboring planar surfaces. Second, four different smart sampling and enhancement choices and a path-direction-based

navigation method are proposed, all using the patch-based 3D and color segmentation results. This paper will focus on the second step.

Fig. 3 shows a real example of the patch-based method. Fig. 3(a) represents a reference image (left view) of a pair of stereo images and Fig 3(b) shows its corresponding patch-based 3D model. The plane parameters in the camera coordinate (3D model) and the boundaries of a few large patches are shown.

In order to make full use of the limited resolution of alternative perception devices such as a retinal prosthesis and tongue stimulator, starting with the patched-based 3D representation, our approach provides a number of different alternative perception choices to end users, including: smart sampling, background removal, motion parallax simulation, image re-illuminating, dynamic highlighting of objects-of-interest, and path directions. The end user could select one or more perception methods. We will discuss each of these methods in the following sub-sections.

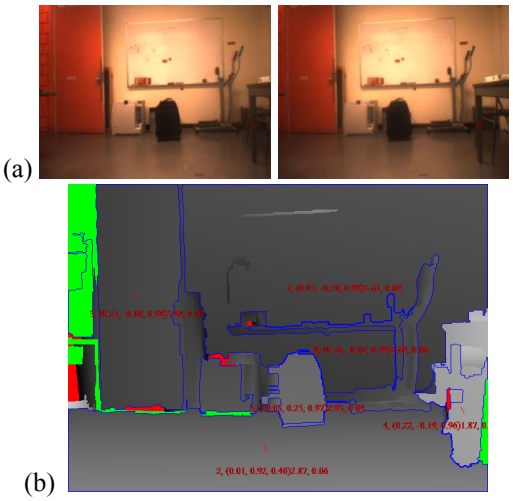

Fig. 3. (a) A stereo pair of color images and (b) the depth map generated by patch-based method (the brighter, the closer). For several large regions indexed, the boundaries of regions are marked by closed curves (blue) and planar parameters are drawn on the regions.

4.1 Smart Sampling Using RGB-D Segmentation

Subsampling needs to be conducted to reduce a RGB-D map from an original high resolution image (Ro) to a low resolution sampling (Rs) for visual implants or tongue stimulation. Regular uniform subsampling methods sample one pixel every N pixels (N=Ro/Rs). For some thin objects (whose dimension are smaller than N), for example, a lamp pole in front of the blind user, it is impossible to preserve the object after subsampling if the width of the pole in the image is smaller than N. This can be very dangerous because the user may bump into the objects right in front of him/her.

The smart subsampling we have proposed can preserve such thin objects, which are determined to be significant by a number of measurements: the distance from the user, the confidence in the 3D measurements, and the shapes of the objects. Currently, we consider thin but long objects that could be vertical poles and horizontal bars. From the patch-based stereo vision method, a depth (D) map consists of many planar patches with known geometric relations. However, the regular sampling method does not make use of the patch information. The goal of smart subsampling is to not lose any important information when the subsampling is performed based on the patch-based 3D representation. Note that we would like to subsample both the color/intensity image (RGB map) and its corresponding depth (D) map.

In order to reduce the computational cost for a portable device implementation, we have developed a very simple algorithm based on the 3D-patch-based stereovision result. For more details of the smart sampling algorithm and analysis of computation complexity, please refer to our paper [17]. Here we only provide a brief overview. The smart sampling is performed on each patch and it is guaranteed that at least one pixel can be sampled regardless of the size of the patch. Initially, the sampled image is filled with the regular sampling method. During the smart sampling process, a sampled 3D value is then filled in by comparing the new 3D values with the existing 3D value and the value of the closer 3D value is kept. In this way, thin objects in close range can still be preserved during sampling. With the same method (based on 3D information), the original color image can be sampled and any important information can be kept as well.

The computation of both initialization and smart sampling is O(wh), where w and h represent width and height of the sampled image (both are 20 in our experiment). In addition, because the process of each patch could be performed separately, it can be easily programmed with a parallel processing method. This is extremely useful for real-time processing, such as during navigation.

4.2 Background Removal Based On Scene Labeling

Based on the 3D patches and geometric relations among connected patches in the patch-based 3D representation, simple object detection techniques are applied and the patches can be labeled into different object categories. Background objects, such as static and no-obstacle objects, can simply be removed from the RGB-D map in order to reduce the complexity of the environment scene when presenting to the user; only the objects of interest, such as persons or obstacles, are “displayed” to the blind user. After removing the background, we only keep a small amount of important information in the final sampled RGB-D map and transduce it to end users through different stimulators, such as the tongue stimulator.

In order to allow blind people to have a stronger ‘feeling’ when observing obstacles and other types of objects of interest around him/her, the objects of interest (OIs) can be highlighted by increasing their intensities, whereas the intensities of less-important objects can be decreased. Thus, the contrast between OIs and non-OIs (i.e., background) is

increased so that the end users are more likely to have a better perception of the OIs. Fig. 4a and 4b show the results of background removal in both the depth and the color maps; note that only the pole and the cubic box are kept. Fig. 4c shows the change in intensity of the box.

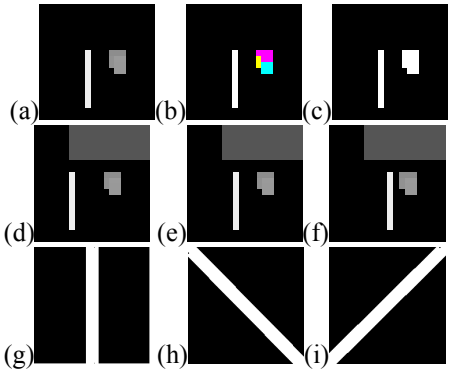

Fig. 4. Background removal, motion parallax, and object re-illumination. (a) The sampled depth image using background removal method. (b) The sampled original image using background removal method (note: the pole and the object are kept in (a) and (b)). (c) The sampled original image using image re-illumination. (d-f) The object of interest (the pole) in close range is shifted left and right to simulate its motion parallax. (g-i) three lines with different orientations represent forward, left turn and right turn navigation directions.

4.3 Motion parallax simulation using 3D

Humans are very good at identifying motion. It has been shown that people can recognize objects mostly by their motion from a video sequence, even with low resolution or largely noise-contaminated images in which the objects cannot be easily identified. Therefore, an alternative way to highlight OIs is to generate motion parallax according to the distance of OI from end users, given that we have obtained the range information of these OIs. Since a closer object produces larger motion parallax, by simulating motion parallax of the objects of interest in a small temporal sequence provided to the user, we hope to provide users a strong perception of the close object. In our experiments (Fig. 4 d-f), an OI (the thin pole) is shifted from left to right, which may produce stronger perceptions than just increasing contrasts between OIs and non-OIs. One limitation of this is that the scene may become too cluttered if there is more than one object being rendered with motion parallax. However, we may consider rendering one object at a time to cope with this limitation.

4.4 Dynamic Object Re-Illumination and Highlighting

By constantly changing the intensities of some OIs, we hope to bring these objects to the attention of the end users. The intensities of an OI can be changed from dark to bright and then changed back recursively. Likewise, we can also highlight small but important objects in close range.

With smart sampling, small objects in close range can be preserved, but they may only occupy 1-2 pixels in the sub-sampled image. In order to provide a clearer stimulus to an end user, the size of the OI can be changed in the rendering step. That is, it can be enlarged (by a morphological operation) until it reaches a certain size and then reduced back to the original size; this process can be performed interactively. Alternatively, we can also highlight the contour of an OI dynamically so that the alternative stimulation can translate the information more efficiently to the user.

4.5 Path Directions

The above smart sampling and enhancement methods are used to preserve objects of interest (OIs) with small size for obstacle avoidance and targets. With RGB-D data, the 3D locations of obstacles are also available. A traversable path can then be provided to the end user in real time; we have found that users are more accurate in discriminating orientations of straight lines using a tongue stimulator (Section 6).

Three or more different line orientation are used to represent forward, left turn and right turn navigation directions (Fig. 4 g-i), We assume the path planning to avoid obstacles and/or to lead to targets are calculated using RGB-D data, we translate the traversable path directions as orientation of lines on the tongue stimulator of a user in real-time to enable smoother and collision-free navigation. Therefore, we can study what information is needed to enable tongue-based navigation using a low resolution visual prosthesis.

5. EXPERIMENTAL AND RESULTS

Experiments have been performed to test our approach in smart sampling. Image sequences were captured by the stereovision head Bumblebee, which is fixed on a mobile platform. For a pair of stereo images, the left camera serves as the reference camera. The baseline distance between the left and the right cameras is 12 cm, and focal length of each is 3.8 mm. The stereo system has been pre-calibrated and image pairs rectified.

Fig. 3a shows a stereo pair of color images captured in an office, and Fig. 3b shows the rendered depth map from the estimated planar representations using the patch-based stereo matching algorithm. The plane parameters (3D model) in the form of "no. (a, b, c) d, n", representing which plane, the planar normal, the average distance of the plane to the viewer and the uncertainty measurement, are marked for a number of large patches. Note that patches with large uncertainties are highlighted in green.

In Fig. 5a, a pair of stereo images of an indoor scene is shown, including a table, a chair, a printer and a tripod, which are about 1 to 4 meters away from the stereovision head. A depth map (the brighter, the closer) rendered from the results of the plane parametric estimation of the patch-based stereo method is shown in Fig. 5b. For several large surfaces, the plane parameters in the form of "no. (a, b, c) d, n" are also shown, with their boundaries highlighted in blue. These plane estimation results are consistent with the results measured by hand. The parametric representation can be

transduced to a blind user easier than an array of depth points with a uniform sampling method.

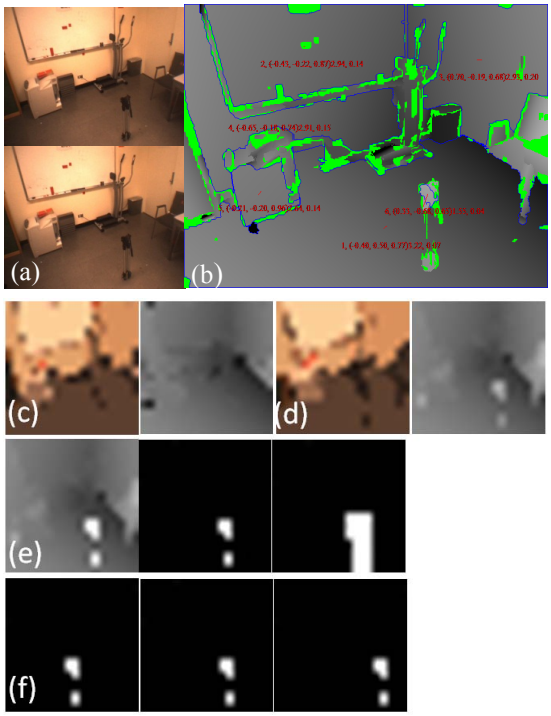

Fig. 5. (a) A pair of stereo images of an indoor scene captured in an office with a number of objects (note: a tripod is in a close range); (b) The depth map of the indoor scene; pixels with large uncertainty are marked in green; (c) sampling results of the RGB and D maps using uniform sampling method: the tripod is missing after regular sampling; (d) sampling results of the RGB and D maps using smart sampling method: the tripod is kept after sampling; (e) sampling results of highlighting objects close to the user, after removing background and applying image dilation; (f) sampling results using motion parallax simulation and only an OI close to the user is shifted towards left, kept in its original position and then shifted towards right, respectively.

Fig. 5 (c-f) shows the results after applying a number of subsampling approaches. Fig. 5c shows that the tripod, which is about 1.55 meters from the user, is missing after uniform sampling, but it is preserved using the proposed smart sub-sampling approach (Fig 5d). In Fig 5e, sampling results are shown after highlighting objects close to the user, removing the background and performing dynamic highlighting, consecutively (from left to right). Three samples of the small motion sequence using the motion parallax simulation are shown in Fig. 5f; only the object of interest close to the user (i.e., the tripod) is shifted left and right to simulate its motion parallax so that users have a better understanding of the 3D distances of OIs.

6. EARLY TESTS USING BRAINPORT

The experiments show that the smart sampling method can be used to automatically select, sample and transduce the most useful scene information to end users through visual substitution devices. However, how effectively blind users could process visual information from the smart sampling is still an open question and it is worth investigating. In this section, we report some preliminary experiments with the BrainPort V100 [1] to show how well users process visual information using such a low-resolution tactile device.

Because the visual environment is typically filled with many different types of objects, the image delivered to the tongue is often cluttered with too much information, making it difficult for the user to process and interpret, especially after down-sampling to the 20×20 electrode array. Additionally, very few studies have objectively assessed the discrimination abilities of users with the BrainPort device. Thus, little is known about the effectiveness of the device in conveying visual information.

In order to examine the baseline abilities of naïve users to discriminate patterns of stimulation delivered to the tongue, we tested subjects that had no prior experience with the device using several simple shape discrimination tasks. It was found that subjects perform with greater accuracy when discriminating between various line orientations than when discriminating between more complex shapes, such as squares and circles. This finding is useful, in that line orientations can be used to encode and convey the directions of traversable paths and maybe also important objects for blind users.

Ongoing work is assessing the effectiveness of these training and preprocessing procedures by measuring obstacle detection and step-by-step navigation. We are creating a program that has a simple maze with a specified destination and an avatar controlled by the blind user from a starting location. The program also computes the shortest path from the starting location to the destination. The program provides the user step by step direction information (to the user's tongue). This experiment serves as a platform for studying what information is needed to enable tongue-based navigation, and to what extent the blind user can move in real environments. With the above experiment design and setup, we can perform different navigation experiments, without blind users walking in real 3D environments. After designing a number 3D mazes, the ground truth (3D map and step by step directions) is also available. Therefore, these experiments can be done with a user sitting in front of a computer.

To test whether continued use and more extensive training improves performance over time, we will conduct another shape discrimination study that incorporates multiple training sessions with the device. We will also test whether dynamic inputs to the tongue stimulator will improve performance. Following the completion of this study, we will then assess whether discrimination ability can be further improved with the use of computer vision preprocessing algorithms that de-clutter and enhance the image before it is sent to the user's tongue. Discrimination performance for images that have been preprocessed with object, edge, and depth detection algorithms will be compared to

discrimination performance for images that have not been preprocessed.

In particular, we will compare how the smart sampling algorithms with both color and depth (RGB-D) segmentation can be used with the BrainPort V100 to better assist a blind user. We will also use the smart sampling algorithms to separate objects of interests, extract their boundaries, and translate them into dynamic forms in order to assess the most effective method for conveying information so that users are able to recognize the objects.

7. SUMMARY AND DISCUSSIONS

In this paper, we present a set of computer vision techniques to improve the performance of visual prostheses. We first perform the patch-based method to generate a dense depth map with region-based representations, and then we apply a number of smart sampling and enhancement methods to transduce the important/highlighted information, and/or remove background information, before presenting to visually impaired people. The patch-based method generates both a surface based RGB and depth (RGB-D) segmentation instead of just 3D point clouds, therefore, it carries more meaningful information and it is easier to convey the information to the visually impaired.

Transducing digital video images into displays of a low resolution device is required for state-of-the-art visual prosthetics. The proposed smart sampling method can preserve close range objects that are significant by a number of measurements: the distance from the user, the confidence in the 3D measurements, and the shapes of the objects. A set of practical sampling and enhancement methods can be used to extract important information and highlight objects of interest in different ways in order to allow an end user to easily understand the environment. Further, we are conducting more experiments with BrainPort to investigate the effectiveness of both recognition and navigation that blind people can perform using such a low-resolution tactile device.

8. ACKNOWLEDGMENTS

This work is supported by National Science Foundation Emerging Frontiers in Research and Innovation Program under Award No. EFRI-1137172, National Science Foundation Award, No. BCS-0843148 and City SEEDs: City College 2011 President Grant for Interdisciplinary Scientific Research Collaborations. The work is also supported by a PSC-CUNY Research Award (Grand Round 44).

9. REFERENCES

[1] BrainPort Vision Technology. http://vision.wicab.com/technology, last visited November 2012

[2] J.D. Loudin, D.M. Simanovskii, K. Vijayraghavan, C.K. Sramek, A.F. Butterwick, P. Huie, G.Y. McLean, and D.V. Palanker. Optoelectronic retinal prosthesis: system design and performance. *Journal Neural Engineering*, 4 (1): S72–S84. 2007

[3] N Barnes. The role of computer vision in prosthetic vision, *Image and Vision Computing*, 20, 478-439, 2012.

[4] X. He, C. Shen, N. Barnes. Face detection and tracking in video to facilitate face recognition in a visual prosthesis. *Annual Meeting of the Association for Research in Vision and Ophthalmology*, Florida, 2011

[5] L. Horne, N. Barnes, C. McCarthy, X. He. Image segmentation for enhancing symbol recognition in prosthetic vision. *Annual International IEEE Engineering in Medicine and Biology Society Conference*, 2012.

[6] C. McCarth, N. Barnes and P. Lieby. Ground surface segmentation for navigation with a low resolution visual prosthesis. *Annual International IEEE Engineering in Medicine and Biology Society Conference*, Boston, 2011.

[7] J. Coughlan, R. Manduchi, H. Shen, Cell phone-based wayfinding for the visually impaired. *1st International Workshop on Mobile Vision*, 2006.

[8] R. Manduchi, J. Coughlan and V. Ivanchenko. Search strategies of visually impaired persons using a camera phone wayfinding system. *ICCHP* 2008.

[9] R. Audette, J. Balthazaar, C. Dunk, and J. Zelek, A stereo-vision system for the visually impaired. *Tech. Rep, Sch. Eng., Univ. Guelph, Guelph, ON, Canada.* 2000-41x-1, 2000.

[10] L. Gonz´alez-Mora, A. Rodr´ıguez-Hern´andez, L. F. Rodr´ıguez-Ramos,L. D´ıaz-Saco, and N. Sosa. Development of a new space perception system for blind people, based on the creation of a virtual acoustic space. *Technical Report*, May 8 2009

[11] X. Lu and R. Manduchi. Detection and localization of curbs and stairways using stereo vision. *IEEE International Conference on Robotics and Automation*, 2005

[12] V. Pradeep, G. Medioni and J. Weiland. Piecewise planar modeling for step detection using stereo vision. *Workshop on Computer Vision Applications for the Visually Impaired*, 2008

[13] S. Se, B. Michael. Vision-based Detection of Staircases, *Asian Conference on Computer Vision*, 2000

[14] S. Se. Zebra-crossing detection for the partially sighted. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2000

[15] H. Tang, Z. Zhu, J. Xiao, Stereovision-based 3D planar surface estimation for wall-climbing robots. *International Conference on Intelligent Robots and Systems*. 2009.

[16] Second Sight, http://2-sight.eu/en/home-en, last visited November 2012.

[17] H. Tang, T. Ro and Z. Zhu. Smart sampling and transducing 3D scenes for the visually impaired. *IEEE International Conference of Multimedia and Expo*, July 2013.