EdgeSonic: Image Feature Sonification for the Visually Impaired

Tsubasa Yoshida
UEC Tokyo
Tokyo, Japan
tsubasa@vogue.is.uec.ac.jp

Kris M. Kitani
UEC Tokyo
Tokyo, Japan
kitani@is.uec.ac.jp

Hideki Koike
UEC Tokyo
Tokyo, Japan
koike@is.uec.ac.jp

Serge Belongie
UCSD
San Diego, CA, USA
sjb@cs.ucsd.edu

Kevin Schlei
UW-Milwaukee
Milwaukee, WI, USA
kevinschlei@gmail.com

ABSTRACT

We propose a framework to aid a visually impaired user to recognize objects in an image by sonifying image edge features and distance-to-edge maps. Visually impaired people usually touch objects to recognize their shape. However, it is difficult to recognize objects printed on flat surfaces or objects that can only be viewed from a distance, solely with our haptic senses. Our ultimate goal is to aid a visually impaired user to recognize basic object shapes, by transposing them to aural information. Our proposed method provides two types of image sonification: (1) local edge gradient sonification and (2) sonification of the distance to the closest image edge. Our method was implemented on a touch-panel mobile device, which allows the user to aurally explore image context by sliding his finger across the image on the touch screen. Preliminary experiments show that the combination of local edge gradient sonification and distance-to-edge sonification are effective for understanding basic line drawings. Furthermore, our tests show a significant improvement in image understanding with the introduction of proper user training.

Categories and Subject Descriptors

H.5.2 [**Information interfaces and presentation**]: User Interfaces-Auditory (non-speech) feedback

General Terms

Human Factors, Design, Experimentation

Keywords

Image sonification, sensory substitution, visually impaired, edge detection

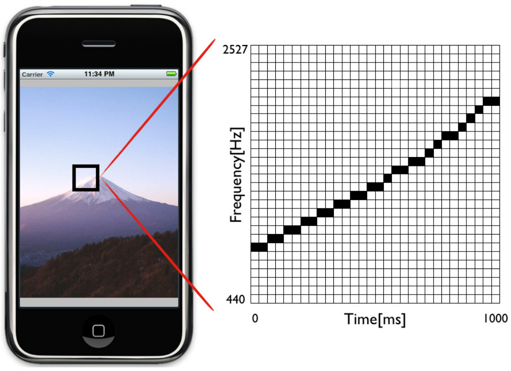


Figure 1: Mapping from image features to sound

1. INTRODUCTION

The visually impaired leverage auditory and haptic senses (among other senses) to recognize the world around them. However, objects displayed on flat surfaces (e.g. posters, digital displays, labels) and objects at a distance (e.g. buildings, landscape, billboards) are harder to perceive. How then can we use technology to translate these types of flat and distant visual information in such a way to aid the visually impaired to perceive them?

Devices such as the OptiCon scanner that displays printed documents on a haptic display, 2D pin arrays and device specific sonification (e.g. digital temperature reader for ovens) have been developed for the visually impaired but can be very costly (e.g. several thousand dollars). Therefore, we aim to develop a framework that is more accessible and affordable to more people.

With the evolution and wide spread use of mobile devices, it is fair to say that many people now have access to a lightweight camera, sufficient computing power and audio playback. In this work, we explore the use of image capture, basic computer vision algorithms and audio feedback supported by existing mobile platforms to aid visually impaired users in accessing visual information.

In our proposed framework we leverage basic image processing techniques to extract salient features from an image and then transpose them into sound. In particular, we extract image edges (regions of high visual contrast) and map them to a combination of timed frequency oscillators (Figure 1). Our prototype system allows a users to take a picture of the visual world and explore the static image aurally via a mobile touch-screen device. Our preliminary tests show that with proper training, our system can be used to understand basic shapes and patterns in under 90 seconds.

2. RELATED WORK

Previous work on image sonification can be roughly divided into two types of sonification. In high-level (symbolic) sonification, visual information is translated into natural language. In contrast, low-level sonification transposes visual information into an abstract audio signal. Our proposed approach falls into the latter category of low-level image sonification.

2.1 High-level sonification

The majority of work on sonification for the blind has focused on high-level (symbolic) sonification. Text-to-speech (TTS) is the most well known sonification system, where such software as the VoiceOver function on Apple products and JAWS (Freedom Scientific, Inc.) can sonify text characters and objects displayed by a computer. The advantage of such systems is that they map visual information to the information-rich realm of the natural language. The obvious limitation of high-level sonification is that it is limited to objects that have obvious semantic representations. For example, it is not clear how to sonify complex shapes, color variations and detailed textures.

LookTel[1] goes beyond TTS and implements computer vision algorithms to automatically recognize object categories and aurally returns the name of an object. Although training the classifiers requires a potentially extensive training process, virtually no effort is required of the users to utilize the system.

The VizWiz [3] mobile phone application allows a visually impaired user to tap into the power of crowdsourcing to obtain answers to visual queries. The system combines the power of TTS and a pool of remote sighted guides to aid the visually impaired user. The main advantage of the system is that it leverages the brain power of humans and therefore can deal with a large range of complex queries. The lag time between queries and answers, the running cost of queries and the availability of remote sighted users is still an open issue.

2.2 Low-level sonification

While high-level sonification eases the burden of recognition, it is also limited by the lexicon of the system. In contrast, a mapping to an abstract audio space (low-level sonification) has the advantage of dealing with a wider range of objects without being constrained by a lexicon. Low-level sonification can still work with hard-to-label (untrained) objects and can work in realtime without relying on remote guides.

The vOICe system [4] sonifies the global luminance of an image and maps luminance values to a mixture of frequency oscillators. Specifically, the image brightness is mapped to amplitude and location is mapped to a frequency. The vOICe system scans the entire field of view of a head mounted camera with a vertical bar from left to right and transposes the luminance over the vertical bar to sound. One of the advantages of the vOICe system is that it sonifies an entire image to convey the global content of any type of scene and the system does not require any type of prior training or lexicon.

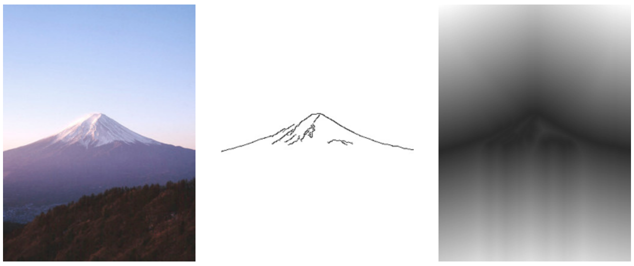


Figure 2: (a)source image (b)edge image (c)distance image

The Timbremap [5] system sonifies local visual information base on the location of a users finger on a map. Timbremap helps the user to navigate through a map by sonifying distances to lines (streets) on the map. By placing his finger on a map, the position of the finger with respect to the nearest line is transposed into an audio signal. The system uses stereo panning to convey whether the finger is to the right or left of the line. Although the Timbremap was designed for maps, the concept of binaural feedback can be applied to any type of line drawing.

Ivan and Radek[2] presented a sonification method for mapping color information to a frequency oscillator, where color information was mapped to the wave envelope, waveform and frequency.

A common attribute of low-level image sonification is that it requires the user to learn the mapping between visual features and audio feedback. While it does place a greater burden on the user, it also taps into the human potential for sensory substitution and uses a diverse set of cognitive skills. We hypothesize that with the proper training, low-level sonification techniques offer users more depth and breadth in analyzing audio-visual information.

3. OUR APPROACH

The use of our fingers is an intuitive way to explore local texture and shape. For example, we can feel the grains on a plank of wood or follow a crease in a piece of paper. For a visually impaired person trained to read Braille, the fingertips are used as a type of local area sensor to understand Braille dot patterns or raised figures. In a similar manner, we aim to extend the analogy of the finger as a local area sensor to provide an intuitive mode of obtaining local spatial information. When a finger touches an edge, local area sonification occurs. From initial experiments, we found that local area sonification was only effective when a finger was located near an edge in the image. When no edge existed in the vicinity of the finger, the user would wander the image and lose his sense of positioning relative to an edge. To aid the user in areas with no edge features, we incorporated a secondary sonification mode, distance-to-edge sonification, that conveys the distance to the nearest edge. We give a detailed explanation of these two modes of sonification in the following sections.

3.1 Sonifying the Edge Image

As mentioned above, we begin our exploration of image sonification with the use of image edge features. In visual perception, is has been shown that people often use the image edges (object outlines) to recognize the object [6] [7]. Therefore, in this work, we explore the use of image edge sonification. To extract edge features from the image, we use the Canny edge detection [8] to obtain contours. The Canny edge detector is relatively robust to noise and two thresholding parameters can be adjusted to extract only the dominant edges in an image (see Figure 2(b)). Although we have utilized the Canny detector for simplicity, we believe that more sophisticated contour detectors such as gPb [10] will improve the quality of the extracted edge map. This edge image is used to generate the audio feedback for local area sonification.

When a finger is placed at a pixel location i such that it is part of an edge $i \in E$, a small vertical bar scans the image directly under the finger (see Figure 1). Each element of the vertical bar is associated to a frequency oscillator (a simple sine wave for our experiments), where each element is mapped on an exponential scale over a range between $f_{max} = 2527Hz$ (top element) and $f_{min} = 440Hz$ (bottom element). As the vertical bar scans the local area in the binary edge image, a certain frequency oscillator is turned on if the pixel that it scans is an edge, otherwise the oscillator is turned off. A horizontal line in the edge image yields long lasting single frequency sine wave. A vertical line yields a single bleep sound, where all frequency oscillators are turned on simultaneously for a short duration.

We implemented our image sonification system on a mobile touch-screen display device, the Apple iPhone 3G. The size of the square area scanned by the local area sonification mode is 30 × 30 pixels, roughly the size of the finger tip. Since edges are usually very thin and very hard to localize with the fingertip, the edge image was dilated and smoothed to increase the width of the edges.

3.2 Sonifying the Distance-to-Edge Map

In addition to the edge image, we calculated the shortest distance to the nearest edge for use with distance-to-edge sonification. Each element j in the distance-to-edge map contains the Euclidean distance to the nearest edge (some pixel location in the image). The resulting distance-to-edge map generated using the Felzenszwalb algorithm [9] is shown in Figure 2(c). The distance map is used to generate a pulse train to convey to the user the distance $d(j)$ to the nearest edge, where the mapping from distance to frequency is given as below.

$$f(j) = (f_H - f_L)\left(\frac{255 - d(j)}{255}\right)^2 + f_L \qquad (1)$$

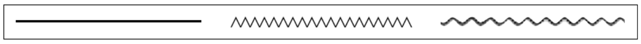
where f_H is the highest frequency of the pulse train and f_L is the lowest output frequency of the pulse train. We normalize by 255 because the maximum value of the distance image has been scaled to 255.

As the user slides his finger closer to an edge, the frequency of the pulse train increases and reaches a maximum when the position is 10 pixels away from an edge. The pulse train ceases to play once the user's finger is within 10 pixels of an edge.

4. LOCAL AREA SONIFICATION

To understand the effect of local area sonification on image understanding, we performed an experiment where a user is asked to reproduce three line drawings shown in Figure 3. In the first part of the experiment, the local area sonification is turned off and the user relies solely on the distance-to-edge sonification to reproduce the line drawing. Then in the second part of the experiment, the user reproduces the line drawings using both the local area sonification and distance-to-edge sonification. The users where only given a simple verbal explanation of the sonification and no training was administered to the participants. Each participant was given 60 seconds to reproduce the line drawing.

In Figure 4 we observe that four out of the six participants were able to identify the locally periodic patterns generated by the sine wave. Notice that the gradients of line drawings of all participants changed after including local area sonification.



(a) straight line (b) saw wave shape (c) sine wave shape

Figure 3: Ground truth for localized patterns

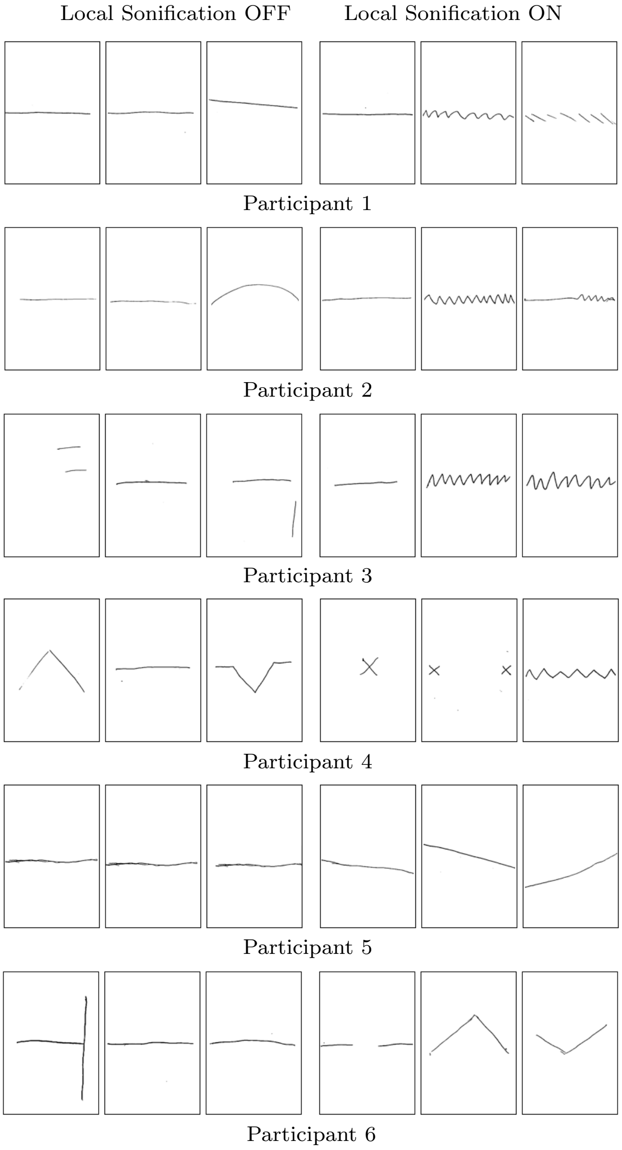


Figure 4: User results: Local area sonification OFF (left) and Local area sonification ON(right)

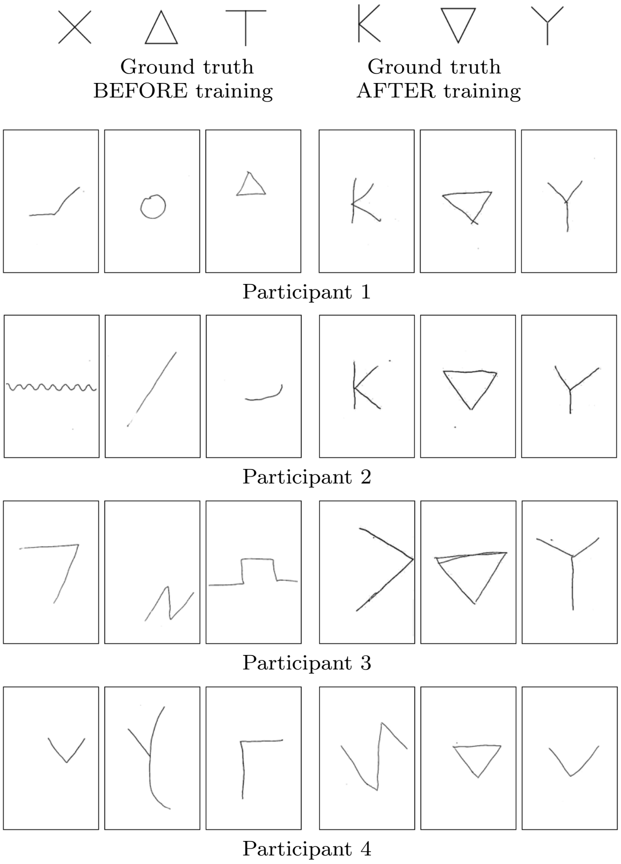


Figure 5: User results: Before training (left), After training (right)

5. USER TRAINING

In this second experiment, we tested the effect of training. The line drawing produced by the participants before and after training are given in Figure 5. Participants were only given 90 seconds to reproduce the line drawings for both tests. Before training, none of the participants were able to reproduce any of the line drawing. For the training stage, each user was given roughly 20 minutes to explore various basic shapes (with prior knowledge of the line drawing). The moderator also quizzed participants to localize slopes, corners and t-junctions. After training, we observe a significant increase in performance. Two of the four participants were able to correctly reproduce all line drawings. All participants were able to reproduce the triangle. We also note that the degree to which the reproductions differ from the ground truth line drawings was also significantly reduced after training. Although some of the reproductions are incomplete, none of the participants generated lines with gradients that contradicted the ground truth line drawings.

6. DISCUSSION AND CONCLUSION

Even with the use of distance-to-edge sonification, many participants commented that it was difficult to track absolute position in the current framework. We would expect that a hybrid sonification scheme using both local and global sonification may help to alleviate this problem.

Although all user tests were performed with blindfolded sighted participants, we are planning to evaluate our system with people with congenital blindness and those who have lost their sight later in life. We noticed during our experiments that participants were highly influenced by their prior knowledge of the visual world and rudimentary shapes (e.g. perceiving a triangle as a circle). It will be interesting to observe how this phenomenon comes into play for the visually impaired.

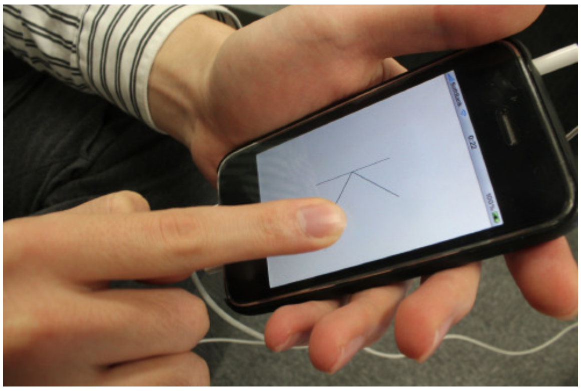


Figure 6: Test participant using the EdgeSonic system

In this paper, we presented a sonification methodology based on edge gradients and distance-to-edge maps. Preliminary experiments with blindfolded sighted persons showed that local area sonification enabled participants to be more sensitive to changes in the local gradients in images. In addition, experiments showed significant improvements in the participant's ability to reproduce the line gradients in line drawings after a period of training. Future work will focus on better training techniques and hybrid sonification schemes to increase recognition speed.

7. REFERENCES

[1] J. Sudol, O. Dialameh, C. Blanchard and T. Dorcey. LookTel: A Comprehensive Platform for Computer-Aided Visual Assistance. In *Proceedings of the Workshop on Computer Vision Applications for the Visually Impaired*, 2008.

[2] K. Ivan and O. Radek. Hybrid Approach to Sonification of Color Images, In *Proceedings of the International Conference on Convergence and Hybrid Information Technology*, 2008.

[3] J.P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R.C. Miller, A. Tatarowicz, B. White, S. White and T. Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the Symposium on User Interface Software and Technology*, 2010.

[4] P.B.L. Meijer. An Experimental System for Auditory Image Representations. In *IEEE Transactions on Biomedical Engineering*, 1993.

[5] J. Su, A. Rosenzweig, A. Goel, E. de Lara and K.N. Truong. Enabling the Visually-Impaired to Use Maps on Touch-Enabled Devices. In *Proceedings of MOBILECHI*, 2010.

[6] R. L. Gregory. Cognitive Contours. In *Nature, vol 238, pp.51-52*, 1972.

[7] I. Rock and R. Anson. Illusory Contours as the Solution to a Problem. In *Perception, vol.8, pp.665-681*, 1979.

[8] J. Canny. A Computational Approach to Edge Detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 1986.

[9] P.F. Felzenszwalb and D.P. Huttenlocher. Distance Transforms of Sampled Functions. In *Cornell Computing and Information Science Technical Report*, 2004.

[10] M. Maire, P. Arbeláez, C. Fowlkes and J. Malik. Using Contours to Detect and Localize Junctions in Natural Images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.