AUDITORY EYES: REPRESENTING VISUAL INFORMATION IN SOUND AND TACTILE CUES

Suresh Matta, Heiko Rudolph, and Dinesh K Kumar

School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology,
410,Elizabeth Street, Melbourne-3000, Australia
phone: + (61) 3 9925 5361, fax: + (61) 3 99255340, email: s2002812@student.rmit.edu.au
web: www.rmit.edu.au

ABSTRACT

This paper presents a theoretical system that provides auditory image representations as an approximate substitute for vision. It can be used for the blind in order to help them navigate. Humans are very sensitive to sound, so expressing the visual environment in terms of audio sensations can support visually impaired people in their day-to-day routines. We postulate imagining sounds will help people with disabilities such as blindness by substituting one sensory mode with others. Our research aims to provide suitable solution to this by integrating two-dimensional image-processing techniques with depth detection and sound. In addition we also propose to use the tactile input channel to convey information to the blind.

1. INTRODUCTION

Human vision requires many low-level capabilities for instance, the ability to extract images of lightness, color and range. An important low-level capability is object perception, a capability of figure/ground discrimination that separates objects from the background. A normal human has ability to do this discrimination but the visually impaired has limited or no ability so current research is developing methods to provide this ability using other modalities.

Humans have very well developed auditory sensing ability. They employ multiple channels (sounds, vision, haptic etc) in their everyday lives to support many of their day-to-day routines. Humans easily identify a wide variety of sounds such as splashes, pouring, streaming, breaking waves and dripping [10]. Exploiting this human capability we can convey many kinds of information to the visually impaired.

It is possible to create an impression of objects or things that a person is looking at but requires the design of complex sounds and selected suitable musical parameters. If the sound scape of events can be conveyed via the auditory channel, it certainly helps the task of understanding events, objects and textures. It is therefore possible to substitute the ears with eyes to represent the physical world to some extent.

This paper is structured as follows. Section 2 presents previous work; section 3 explains the proposed theory for 'Auditory Eyes' design and an overview, section 4 presents the proposed method of image to audio translation and section 5 & 6 presents a brief idea of sound generation and tactile generation methods. Finally the last section presents the conclusion.

2. RELATED WORK

An example of data to sound mapping is [11], where streams of images are transformed into a time-multiplexed auditory representation is outlined below. The new frame (image) is sampled, digitized and stored as an M (height/ rows) × N (width/ columns) pixel matrix. Every row and column is individually averaged and the mapping translates vertical position into frequency and horizontal position into time delay and the brightness represented by the amplitude. A click is generated to mark the beginning of the new frame, or, equivalently, the ending of the previous image. The system requires extensive training of the user.

The sound patterns corresponding to simple shapes are easily imagined. More realistic images yield more complicated sound patterns and learning more complicated patterns can be difficult because of the simplicity of the mapping algorithm. This image-to-sound mapping does not exploit the ability of the human hearing system to detect time delays.

Prior knowledge of the sound is required to use this system and is difficult to recognise the new sounds, and hence training is required. User has to remember the different sounds to understand the different objects and there is a scope for confusion when two sounds are nearly similar.

Another important work reported in this area is the modern Optophone [2], where the image is directly scanned and digitized. The image is then broken down into vertical strips, with high frequency musical notes being generated for pixels located at the top of the strip and low frequency notes at the bottom. It is intuitively natural for the auditory system to associate higher frequencies with a higher vertical position in space [2]. This instantly aids recognition and comprehension of the sounds. The amplitude of the sounds generated varies proportionally to the image intensity.

The difficulty of this technique is that the while human hearing has dynamic capabilities, it is unable to compete with the eye for extracting the static details of the image. Thus this technique is unable to provide detailed information to the user.

3. SYSTEM DESIGN AND OVERVIEW

The conceptual model of Auditory Eyes is illustrated in Fig. 1. The image encoder (Edge and Distance detector) captures images from the cameras and stores them as an N (rows) * M (columns) pixel matrix, separates all objects from their background and measures their distance from the user, the Visual-to-Audio/Tactile Mapper maps visual information into sounds and tactile vibrations, the Sound Generator generates audio impressions of the visual scene and Tactile Generator creates tactile sense.

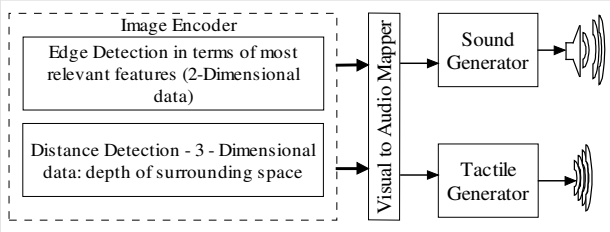

Figure 1: System Overview

3.1 Image Encoder

Image encoder uses edge detection method to separate objects from the background of the visual scene. Subsequently these objects translated into auditory/tactile perceptions. For effective distance estimation triangulation camera setup is used.

The image encoding procedure starts with selecting each digital form of an image among sequence of images and storing it as an N (rows) by M (columns) matrix. Each image will be segmented to identify different objects in an image by identifying the edges of the objects using Sobel edge detection operator. An edge will have higher pixel intensity values than those surrounding it. An edge is declared if the value of the gradient exceeds specified threshold. A 3 X 3 convolution mask K is applied to each pixel p, with horizontal and vertical orientations as shown in the equation (1).

$$O(i,j) = \sum_{k=1}^{n} \sum_{j=1}^{m} K(k,l)\, p(i+k-1, j+l-1) \qquad (1)$$

Where i=1…. N-n+1 and j=1…. M-m+1

Where n and m are rows and columns of convolution mask. In each orientation a gradient component called G_x and G_y will be created. Combining ($|G| = \sqrt{G_x^2 + G_y^2}$) together the absolute magnitude of the gradient at each point is found out. If the magnitude is greater than the threshold, the pixel is marked in black as an edge. Otherwise, the pixel is set to white.

Figure2 demonstrates the distance estimation method. Two cameras will be positioned separating apart a small distance to take pictures from two perspectives. Distance Encoder captures the two images and grabs the pixels using image grabber and calculates the object's depth using matching points in the images in the following way. Two cameras are positioned apart by baseline distance. The '*f*' is the focal length of the two cameras and 'T' is the separation distance between them.

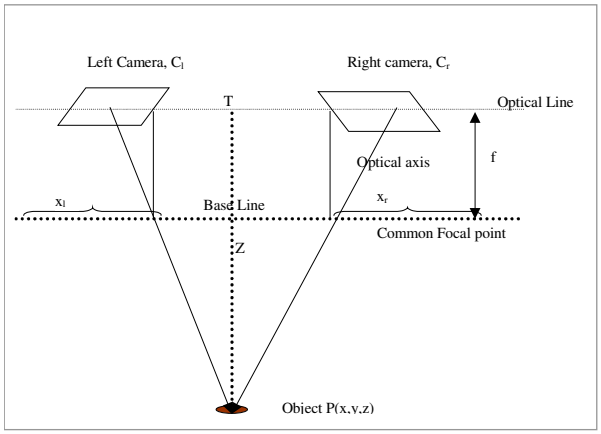

Figure: 2 Depth Calculations

Also from the Figure4, this can be written in the form of an equation (2).

$$\frac{Z}{Z-f} = \frac{T}{T - x_l + x_r} = \frac{T}{T-d} \qquad (2)$$

Where disparity, $d = x_l - x_r$.

Based on above equation the distance of an object from a camera position can be defined as $Z = fT/d$.

4. IMAGE TO AUDIO TRANSLATION

In image to sound transformation three main acoustic factors are amplitude (loudness), pitch and timbre, which enable humans to distinguish the various sounds. This mapping process translates image attributes such as size, brightness, distance, elevation, and azimuth in to musical parameters such as timbre, density, intensity, timing, pitch, and so on. The Table1 shows the proposed mapping of image to sound for Auditory Eyes.

Visual Space	***Audio Space***
Motion	Frequency Shift + Interaural Time Difference +Inverse Square Law
Brightness	Pitch
Space (Distance, Azimuth, Elevation)	HRTF (Amplitude, Reverberation, Azimuth, Elevation)
Edge	Duration

Table1: Image to Sound Mapping

4.1 Representation of Motion in Sound

Motion is calculated from feature matches between images those captured using a pair of cameras. The Figure3 shows the schematic procedure for motion representation in sound. Estimated motion is represented using Doppler effect, interaural time difference and inverse square law. These three different parameters will be simulated in Head related transfer functions, which give the musical effect about the

distance between the object and the listener and motion of the object that passes the listener.

The intensity of the sound varies as a function of distance, and that the frequency and interaural phase of the sound vary as a function of relative velocity with respect to the receiver [9]. The shift in frequency that results from a source moving with respect to the medium is referred to as the Doppler effect. Interaural Time Difference (ITD) is the difference between the time of arrival of sound at each ear caused by the reflection of and diffraction of the head [7].

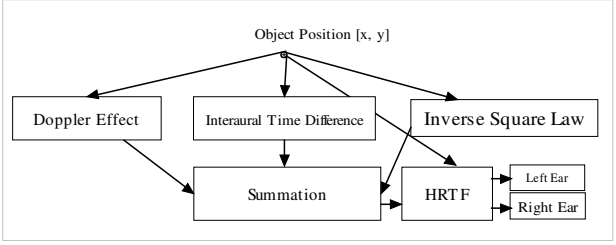

Figure 3: Motion Synthesis

4.2 Representation of Space in Sound

Representation of an object's position in space using sound is possible by using these factors: azimuth, elevation and range. Interaural Intensity and Time differences (IID and ITD) [12] are used for estimating the apparent direction of a sound source. Space can be synthesised using head related transfer functions by simulating these three parameters. These three parameters will be discussed in the sections 4.2.1 and 4.2.2.

4.2.1 Distance Perception

Distance perception will be created to the blind using intensity changes and reverberation effect. The distance estimation effects are observed in [13].

4.2.1.1 Loudness for Distance perception

Inverse Square Law [1] states that intensity will fall exactly by 6dB every time omni directional sound source distance doubles. The loudness increment can be calculated relative to distance using inverse square law as stated in equation (3):

$$db = 20 \log_{10} [1/\text{Distance Increment}] \quad (3)$$

Intensity level at 0dB gives the perception of nearest object and when the object goes far away, intensity decreases. Amplitude manipulations alone do not always give a sense of distance, but the effects of amplitude are stronger for unfamiliar sounds than they are for familiar ones [6].

4.2.1.2 Reverberation for Distance Perception

As reverberation for distance cue it is possible to perceive foreground objects separated from the background. The loudness of the reverberation relative to the loudness of the foreground sound is an important distance cue. The ratio of the direct to reverberant amplitude is greater with nearby objects than it is with distant objects. Reverberation has influence on the properties of the surrounding environment and the perceived spaciousness of a room increases with reverberation time, reverberation level, and/or the amount of decorrelation between left and right ear signals [14].

4.2.2 Azimuth and Elevation

Azimuth is the "horizontal direction expressed as the angular distance between the direction of a fixed point (as the observer's heading) and the direction of the object". This is an angular distance along the horizontal plane to the location of the object. As shown in the Figure4, for objects which are in front of the user will have azimuth=0°, for left side objects -45° and for right side objects 45°.

Elevation is an angle measured from the closest point on the horizontal plane to the above horizontal plane where the object is located. 0 degrees is parallel to horizontal plane while 90 degrees is straight up from horizontal plane.

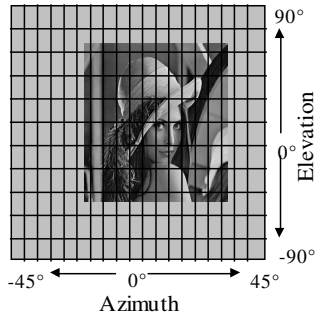

Figure 4: Azimuth, Elevation Measurement relative to the image

Figure4 shows the proposed method of estimation of azimuth and elevation of the object from two-dimensional view. The image is digitized and displays as a N (rows) by M (columns) pixel matrix window. The image window is divided into the X-axis, represented by Azimuth and Y-axis, represented by Elevation. This technique mainly concentrates on the area directly in front of the person, which is azimuth 0°, the centre of the window. -45° is considered as left side and 45° as right side of the user. Elevation 0° to 90°is considered as above the horizontal plane and 0° to -90° as below the camera position.

4.3 Pixel Brightness

Frequency of a sound wave is referring to the rate at which the signal changes with time. Similarly frequencies in an image are referring to changes occurring in space. It is possible to determine the frequencies present in images using Fourier theory [4]. So frequency is estimated from the variations in brightness levels for instance, rapid variations

in brightness level as high frequency and low frequency from smooth areas with little variation in brightness level. The frequency values vary as the intensity varies from dark pixel to light pixel.

4.4 Image Edges

Edges in images are known to be very important for human perception [15]. Humans recognize the size of the objects by their edges. So the proposed model computes the duration (t) of the musical tone from the maximum distance between two points on the edges of the object. Here the time function runs from 0 to t for each object and then starts over for the next object. Suppose there are 'n' objects to perform, if one object takes 't_i' units of time, then all objects processed sequentially then time to process all objects in a scene is $\sum_{i=1}^{n} t_i$. The total time to process 'n' objects in a scene is $T = Log_2 n$.

5. SOUND GENERATOR

For mapping visual information to sounds Csound music synthesis tool is utilized. To create auditory representations, score and orchestra files will be generated using a series of identified parameters from the image to audio mapper. The score file contains the information of different events and notes of music and orchestra file contains the various instruments information. Executing user-defined instruments from orchestra file and by interpreting note events and parameter data from the score file creates audio files.

6. TACTILE GENERATOR

Presenting information in audio is temporal so using tactile generator we can produce cutaneous perception, which will give the user, information about sharp edges, potholes, etc. These cutaneous messages can be created using various parameters such as timbre, frequency, amplitude etc. Frequency range between 10-500 Hz will be used for haptic sensations but maximum sensitivity occurs around 250Hz [8]. Amplitude will be used to express the intensity of stimulation and this range should not exceed 55dB [8], above this range it is painful. Different durations are used to represent different stimuli.

7. CONCLUSION

This research aims to contribute to work in using the aural sense to convey visual information. The visual system of humans is difficult to replace because of its higher information processing ability. Auditory and tactile channels are the next highest in their ability to carry information and can be used to provide partial replacement of sight.

A theoretical system for mapping visual information to auditory and tactile displays has been proposed. Feature selection, or edge detection is employed in order to separate objects from their background; depth is determined using triangulation methods. Relative motion and a subjective sense of depth and space are generated through the use of Head Related Transfer Functions and the Doppler effect.

The work proposes a system of providing information of real world objects via meaningful sounds to the blind, by using audio and tactile information processing capabilities.

REFERENCES

[1] D. R. Begault, *3-D Sound: For Virtual Reality and Multimedia*. MA: AP Professional Publishers, pp. 293, (1994).

[2] M. Capp and P. Picton, "The Optophone: an Electronic Blind Aid," The Engineering Science and Education Journal,(2000).

[3] C. I. Cheng and G. H. Wakefield, "Moving Sound Source Synthesis for Binaural Electroacoustic Music Using Interpolated Head-Related Transfer Functions (HRTFs)," Computer Music Journal, vol. 25, pp. 57–80, (2001).

[4] N. Efford, *Digital Image Processing: A practical introduction using Java,* USA: Pearson Education Limited,(2000).

[5] W. G. Gardner, *3D Audio and Acoustic Environment Modeling*, Wave Arts Inc,(1999).

[6] W. W. Gaver, *Auditory Interfaces,* in Handbook of Human-Computer Interaction. Amsterdam, The Netherlands: Elsevier Science B.V, pp. 1003-1042, (1997).

[7] W. L. Gulick, *Hearing: Physiology and Psychophysics*. London: Oxford University Press,(1971).

[8] E. Gunther, G. Davenport and S. O'Modhrain, "Cutaneous Grooves: Composing for the Sense of Touch," in Proceedings of Conference on New Instruments for Musical Expression, vol. 1. Dublin, Ireland, pp. 6,(2002).

[9] R. L. Jenison, *On Acoustic Information for Motion*, Ecological Psychology, vol. 9, pp. 131-151,(1997).

[10] D. Keesvanden, "Physically-Based Models for Liquid sounds," in Proceedings of the International Conference on Auditory Display, (2004).

[11] P. B. L. Meijer, "An Experimental System for Auditory Image Representations," in IEEE Transactions on Biomedical Engineering, vol. 39, pp. 112-121, (1992).

[12] D. Rocchesso, *Introduction to Sound Processing*,(2003).

[13] B. G. Shin-Cunningham, "Distance cues for virtual auditory space," in Proceedings of the First IEEE Pacific-Rim Conference on Multimedia. Sydney, Australia, (2000).

[14] B. Shinn-Cunningham, "Learning Reverberation: Considerations for Spatial Auditory Displays," in Proceedings of the 2000 International Conference on Auditory Display. Atlanta, GA, (2000).

[15] P. W. Wong and S. Noyes, "Space-Frequency Localized Image Compression," IEEE Transactions on Image Processing, vol. 3, pp. 302-307, (1994).