Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, July 6-9, 2003

SOUNDVIEW: SENSING COLOR IMAGES BY KINESTHETIC AUDIO

Kees van den Doel

Department of Computer Science
University of British Columbia
Vancouver, Canada
kvdoel@cs.ubc.ca

ABSTRACT

An experimental system called SoundView has been developed, which allows the exploration of a color image through touch and hearing. The image is mapped onto a virtual surface with a fine-grained color dependent roughness texture. The user explores the image by moving a pointer device over the image. The pointer acts like a virtual gramophone needle, and the sound produced depends on the motion as well as on the color of the area explored. The roughness texture is obtained by constructing a mapping of three dimensional color space onto a three dimensional sound space. The mapping tries to achieve maximal alignment of the color and sound spaces by preserving the perceptual metrical and topological structure of color space, as well as by incorporating common associations between sound and color.

1. INTRODUCTION

The human vision system and the auditory system are very different, which makes it difficult to present visual information through audio or vice versa. Most attempts at making visual information accessible through sound are very domain specific. For example, a GUI can be made accessible without vision by associating specific sounds with specific GUI elements [1]. Though useful in its domain, this does not offer a method to make generic visual information accessible in situations where vision is not available, either because the user is visually impaired or due to other circumstances (for example when driving a car).

The difficulty with sonifying a generic image is that sound is physically temporal and zero dimensional, a scalar function of time, whereas an image is a two dimensional collection of colors, and non temporal. How can we represent a two dimensional surface of colors in space by a single function of time? The answer may lie in the human sensory system. A sound is not perceived as a rapidly changing pressure but has perceptual dimensions such as spatial location, timbre, and loudness. An image is not perceived as a surface of colors but segmented into objects, regions, etc. At the neurological level, all sensory inputs are eventually represented by activities of neurons and at this level it seems plausible that information can be transferred from one sense to another. This is indeed what happens to people who experience synesthesia [2]. Synesthesia is the involuntary physical experience of a cross-modal association, i.e., a crossing of the senses. Having one sense stimulated causes a stimulation in another sense as well.

In this paper we will describe the SoundView system which allows a user to sense a static image synesthetically through sound and touch. This is done by constructing a virtual surface with a roughness texture corresponding to the image. Instead of feeling the roughness through touch, the user scrapes the surface with a virtual gramophone needle, which is moved with a pointing device such as a graphics pen, a Phantom force feedback device, or a mouse. The disadvantage of a mouse is that there is no unique mapping between mouse position and pointer position, which may however be offset by its widespread use. The scraping sounds depend on the roughness of the area touched and on the manner of touching, to create the illusion of interacting with a physical object. It is hoped that information from the image can be reconstructed by exploring the virtual surface in this manner, similar to the way a blind person may construct a representation of a persons face by touch. The roughness map is constructed specifically to enable it to create informative sounds when interacted with, and can be seen as a sound imprint on a two dimensional gramophone record. SoundView is made available with source code as part of the JASS SDK [3].

The remainder of this paper is organized as follows. In Section 2 related work is reviewed. In Section 3 the construction of the color to sound map is presented. The design principles are explained first and the implementation is described in Section 4. Conclusions and directions for future work are presented in Section 5.

2. RELATED WORK

Peter Meijer has created a vision substitute for the blind [4, 5, 6], The vOICe, which translates images from a camera on-the-fly into corresponding sounds. This is done by sweeping a vertical scan line periodically over the image. The scan line generates sounds depending on the brightness of the pixels it is crossing and the height is mapped to pitch. Though the sounds thus created are not easily interpreted at first it is hoped that the brain can learn to map the information in the sounds to images, either through induced synesthesia, or simply by providing equivalent information through the auditory channel. In [7] acquired synesthesia was reported to appear in a patient several years after vision loss. The patient experienced visual sensations evoked by tactile stimuli on the hands.

In [8] a binaural sensor modeled on the bat sonar system is described. The system allows blind users to determine the spatial location of objects.

Various attempts have been made to make visual information available through haptic devices. In [9] a haptic device for the display of 3D objects and textures was described and user studies on blind and sighted people were performed to determine their ability to determine object properties such as size, angles, and roughness. Complex object recognition was also investigated. User studies in-

vestigating the ability of blind people to use a haptic device to perform various task were presented in [10]. The TACTICS system described in [11] allows the printing of complex images as tactile maps on microcapsule paper. It was found that preprocessing the images by edge detection and enhancement resulted in greatly improved performance in recognition tasks. Attempts to augment the haptic display with auditory information are described in [12], where line graphs are displayed through a combination of haptics and sound. Multimodal perception of roughness textures through sound and haptics is described in [13]. Roughness is displayed aurally by piano tones of various frequencies.

3. DESIGN OF THE COLOR TO SOUND MAPPING

The sonification of an image starts by constructing a virtual surface roughness texture map by mapping different colors to different surface roughness textures. We can think of this stage of the design, which involves no sound yet, as imprinting a two dimensional phonograph with a texture for later playback. Sound will only be produced when a user probes the image with a pointer device which moves a virtual gramophone needle over the surface. The sound that is heard will thus depend both on the color of the image at a particular location as well as on the speed of the motion. We shall refer to "the sound" corresponding to a specific color when we mean the sound produced when the needle is moved with some standard velocity.

The color of a pixel is represented in HSB space by its hue, saturation, and brightness. This three dimensional color space is mapped onto a three dimensional sound space. The sound corresponding to a particular color is interpreted as a roughness texture played back at unit velocity. We have tried to preserve as much of the metrical and topological perceptual structure of the color space as possible, and tried to incorporate color information as a "refinement", rather than as an essential feature of the sound, just as a color picture can be seen as a refinement of a grayscale image, but not something essentially different. We have also tried to incorporate common associations between sound and color such as between white noise and the color white, and the association between color and pitched sounds. These constraints rule out naive mappings such as associating specific tones with the red, green, and blue components of a color.

The sound for grayscale colors (saturation $s = 0$) is generated by filtering a white noise source with a lowpass filter with a cutoff frequency f_c that depends on the brightness, with black ($b = 0$) having the lowest f_c and white the largest f_c.

Color is added by adding a second filter after the lowpass filter, see Figure 1. This filter is parametrized by the hue h and the saturation s of the color. We know that for $s = 0$ the filter should not have any effect on the sound, i.e., have a flat frequency response, in order to reduce to the grayscale model. A natural choice is a reson filter whose spectrum has a single peak with center frequency f_0 and width d_0. The width of the peak is related to the saturation of the color. When we make d_0 large enough the peak becomes so wide it passes all frequencies, suitable for $s = 0$. If the peak is very narrow, the sound will be perceived as pitched. It seems natural to map the frequency f_0 to the hue of the color, but this does not take into account the circular nature of hue, and would cause a discontinuity in sound for nearby colors that cross the $h = 1$ boundary. This can be taken into account by duplication the reson filter with copies at octaves apart, both up and down to slightly beyond the audible range. This creates a filter which generates

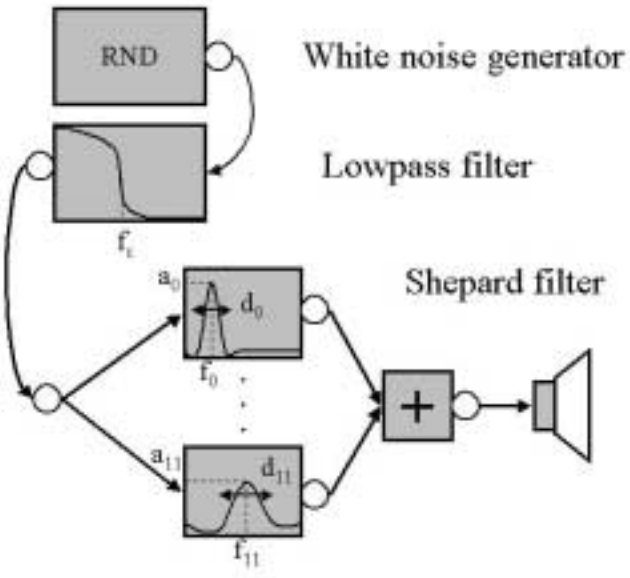

Figure 1: Filter graph to generate color dependent sounds. White noise is filtered through a lowpass filter, with brightness and velocity dependent cutoff frequency f_c. The result is filtered through a bank of 12 parallel reson filters at octaves apart, a "Shepard filter". The resonance frequencies, widths, and gains are dependent on the color and slide velocity.

Shepard tones [14] for small values of d. If all reson frequencies are translated up by an octave the original sound is recovered. It seems therefore natural to map hue $0 \leq h \leq 1$ to an octave range of a "Shepard filter". The gains of the reson filters will be chosen such that the perceived loudness of the sound does not significantly depend on the hue and saturation.

The user interacts with the image by moving a pointer over it, for example with a mouse or (better) a graphics pen or (even better) a force feedback device. The normalized pointer velocity is a dimensionless quantity defined as $v = v_{\text{physical}}/v_0$, where v_{physical} is the actual pointer velocity in m/s, and v_0 is a "normal" velocity with which the user expects to interact. Changing v_0 allows the user to probe large or small features in the image by moving the pointer rapidly or slowly and plays the role of a magnification factor. We allow values of v up to $v_+ = 2.5$ and down to $v_- = 0.01$, lower and higher values are truncated to those limits.

We have tried to map the color parameters to the sound parameters in a manner that preserves the "feel" of color changes, so that equal changes in color correspond to equal changes in sound. The perceptual change of quantities when parameters are varied has been studied extensively, however we are not aware of any studies that have measured precisely the data (e.g. Stevens law exponents [15] of various quantities) needed to make the mapping as "linear" as possible. We therefore have designed the mapping by ear, using our best subjective judgment.

For this purpose a design program, see Figure 2, was created which displays a window with a single color, and renders the sound corresponding to this color when a user would move a needle with a given velocity over it. The colors and the velocity are set with sliders and various design parameters can also be set with sliders so that we can experiment and try out various settings.

4. IMPLEMENTATION OF THE MAPPING

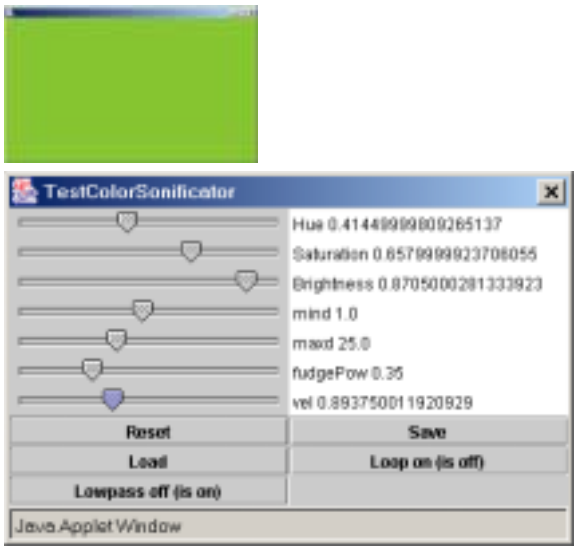

Figure 2: Design applet to tune various parameters that define the color to sound mapping interactively. The Applet can be run in Java 2 enabled web browsers from http://www.cs.ubc.ca/~kvdoel/icad03.

The lowpass cutoff frequency f_c depends on the brightness b and normalized velocity v as follows:

$$f_c = f_-(\frac{f_+}{f_-})^b v,$$

with $f_- \leq f_c \leq f_+$. We found a range of $[f_- - f_+] = [50 - 4000]Hz$ most comfortable and pleasing. Since the human auditory system perceives the logarithm of the frequency as pitch, the pitch of the cutoff frequency appears to move linearly with the brightness b. White is represented by the brightest noise, and black by very low frequency noise close to the lower limit of hearing, which seems very intuitive, at least to people familiar with the term "white noise". The velocity dependence is determined by physical principles from the gramophone needle model.

The color filter consists of a bank of reson filters. The transfer function for a reson filter is given by [16] $\mathcal{H}(z) = 1/(1 - 2R\cos\theta z^{-1} + R^2 z^{-2})$, with $R = e^{-d/F_s}$, $\theta = 2\pi f/F_s$, and F_s the audio sampling rate in Hertz. f is the center frequency and d is the damping factor of the reson. The frequencies for the reson filters are chosen to be

$$f_k = 2^{k+P(h)} v \times 10Hz,$$

for $k = 0, \ldots, 11$, with h the hue. Note that for $0 \leq h \leq 1$ and $v = 1$ the frequencies cover the entire audible range and slightly beyond. $P(h)$ is a non-linear monotonically increasing function of h subject to $P(0) = 0$ and $P(1) = 1$ such that equal changes in P are perceived as equal changes in color. Hue itself is perceptually non-linear as can be seen by pressing the "loop" button in the applet shown in Figure 2. This increments h at a constant rate (identifying $h = 1$ and $h = 0$) but it is apparent that the perceptual rate of change of color is not constant at all.

In order to create as "natural" a sound as possible we choose the damping factors d_k to be proportional to the frequencies for a fixed color and velocity. Such proportionality occurs in a simple model of material elasticity [17, 18] and was shown [19] to evoke the sensation of different materials. The only reason for this choice is to guarantee the generation of more or less familiar sounds, it has no relation to any physical or virtual material property. Thus the dampings are given by $d_k = cf_k$, with

$$c = d_+(\frac{d_-}{d_+})^s / v,$$

where d_- and d_+ are the minimum damping (for fully saturated colors with $s = 1$) and maximum damping (for grayscale colors with $s = 0$). We found the range $[d_- - d_+] = [1-25]s^{-1}$ the most satisfactory. $d = 25/s$ sounds like pure noise, and decreasing d transforms this into a pitched sound until at $d = 1/s$ it has completely lost its noisy character. Increasing d beyond $d = 25/s$ does not produce much audible change and decreasing it below $d = 1$ neither. The factor $1/v$ was added to prevent the reson filters from ringing too long at very low scrape velocity (as the frequencies f_k scale with v).

Each reson also has a gain a_k, which we tried to choose such that the loudness of a sound is independent of h and s. We allow the sound to get softer for lower values of the brightness b as this is a natural side-effect from removing higher frequencies from the sound. The power in a reson excited by white noise is proportional to a_k^2/d_k [20], and since $d_k = cf_k$, if we choose

$$a_k = (d_k)^\sigma = (cf_k)^\sigma = [2^{k+h} d_+(\frac{d_-}{d_+})^s \times 10]^\sigma, \qquad (1)$$

with $\sigma = 1/2$, the power will be independent of saturation and hue. In practice however colors with low vales of s sound louder, using this scheme. This is probably because lower saturation colors generate a broader spectrum an thus excite more critical bands, which is being perceived as louder. Changing the "fudge factor" to $\sigma = 0.35$ produces approximately equal loudness for all values of h and s.

The SoundView system has been implemented in JASS [21, 3], using a Wacom graphics tablet as input device. Because of noise in the pointer position data a lowpass filter was used to smooth the pointer motion, which is sampled at a fixed rate of $44100Hz/64 = 689Hz$. Each time the pen position is read an audio buffer of 64 samples is processed through the filter graph depicted in Figure 1, after computing and setting the filter variables according to the color of the pixel at the pointer and the velocity. The application reads a digital image and the user then interacts with it as described above. A control panel allows the setting of the reference interaction velocity v_0 through a slider. A second slider allows adjustment of the amount of smoothing of the pointer motion, which depends strongly on the type of device used. After the system has been tuned, the control panel can be discarded and a purely auditory interface is presented. In order to evoke an illusion of touching, low latency audio rendering is crucial. This is achieved by using native libraries to access operating system specific audio resources. The system currently runs on Windows, LINUX, and MAC OS/X.

5. RESULTS AND CONCLUSIONS

We have presented the design of SoundView, a system to sense color images by exploring an image with a pointer and listening to

virtual scraping sounds. The color to roughness to sound mapping was designed to maximally align the color and sound spaces by preserving metrical and topological perceptual properties in the map, and by incorporating common associations between color and sound. The complete synthesis algorithm is parametrized by 4 real-time control variables, the color (h, s, b) and the scrape velocity v, which are obtained from the interaction of the user with the image. These four variables are extracted from the user's interaction through the pointer device with the image. The "standard velocity" v_0 which sets the scale for the interaction with the image is intended to be used interactively and consciously by the user. The "design parameters" of the synthesis algorithm have been tweaked to obtain the best correspondence between sound and color as far as we could tell.

Although SoundView was designed primarily as a vision substitute for the blind, various other usages can be imagined such as augmenting feedback in interactions with virtual objects, for example in virtual surgery applications. Clearly user studies are needed to assess the usability of the system. Can people detect basic shapes with the system? How much detail can be perceived in this manner? How does performance on recognition tasks depend on training? Is there significant difference between blind and sighted people? These are important questions which we hope to answer in future work. Currently we are performing user studies in which we try to ascertain the users ability to sense simple shapes using the system.

A natural enhancement of the system would be the integration with haptic feedback, which may seem to be a more natural sensory channel to display roughness textures than audio anyways. Unfortunately, the resolution of force feedback devices does not allow for the display of very fine-grained textures such as described here. However, a haptic device could provide important complementatry information where the audio channel is less informative, especially edge detection and coarse features.

6. REFERENCES

[1] W. W. Gaver, "Synthesizing Auditory Icons," in *Proceedings of the ACM INTERCHI 1993*, 1993, pp. 228–235.

[2] Richard E. Cytowic, *SYNESTHESIA, A Union of the Senses*, The MIT Press, Cambridge, Massachusetts, 2002.

[3] Kees van den Doel, "JASS Website, http://www.cs.ubc.ca/~kvdoel/jass," 2003.

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

[5] Peter Meijer, "The vOICe - Seeing with Sound, http://www.visualprosthesis.com," 2003.

[6] P. B. L. Meijer, "Seeing with Sound for the Blind: Is it Vision?," in *invited presentation at the Tucson 2002 conference on Consciousness, April 8-12, 2002. Abstract no. 187 in "Toward a Science of Consciousness" , in Consciousness Research Abstracts (a service from the Journal of Consciousness Studies)*, Tucson, Arizona, USA, 2002, p. 83, Oxford University Press.

[7] K. C. Armel and V.S. Ramachandran, "Acquired synesthesia in retinitis pigmentosa," *Neurocase*, vol. 5, pp. 2993 – 296, 1999.

[8] Leslie Kay, "A CTFM acoustic spatial sensing technology: its use by blind persons and robots," *Sensor Review*, vol. 19, no. 3, pp. 195–201, 1999.

[9] C. Colwell, H. Petrie, and D. Kornbrot, "Use of a haptic device by blind and sighted people: perception of virtual textures and objects," in *I. Placencia-Porrero and E. Ballabio (Eds.), Improving the quality of life for the European citizen: technology for inclusive design and equality: http://phoenix.herts.ac.uk/SDRU/pubs/TIDE/colwell2.html*, IOS Press, Amsterdam, The Netherlands, 1998.

[10] Calle Sjöström, "Designing Haptic Computer Interfaces for Blind People," in *Sixth International Symposium on Signal Processing and its Applications*, Kuala-Lumpur, Malaysia, 2001.

[11] J. P. Fritz, T. P. Way, and K. E. Barner, "Haptic Representation of Scientific Data for Visually Impaired or Blind Persons," in *Proceedings of the Eleventh Annual Technology and Persons with Disabilities Conference*, California State University, Northridge, Los Angeles, 1996.

[12] R. Ramloll, W. Yu, S. Brewster, B. Riedel, M. Burton, and G. Dimigen, "Constructing sonified haptic line graphs for the blind student: first steps," in *The fourth international ACM conference on Assistive technologies*, Arlington, VA, 2000.

[13] M.R. McGee, P.D. Gray, and S.A. Brewster, "Feeling Rough: Multimodal Perception of Virtual Roughness," in *Proceedings of Eurohaptics*, Birmingham, UK, 2001.

[14] R.N. Shepard, "Circularity in judgments of relative pitch," *J. Acoust. Soc. Am.*, vol. 36, pp. 2346–2353, 1964.

[15] J. G. Snodgrass, Levy-Berger, and M. Haydon, *Human Experimental Psychology*, Oxford University Press, New York, 1985.

[16] K. Steiglitz, *A Digital Signal Processing Primer with Applications to Digital Audio and Computer Music*, Addison-Wesley, New York, 1996.

[17] Richard P. Wildes and Whitman A. Richards, "Recovering material properties from sound," in *Natural Computation*, Whitman Richards, Ed., Cambridge, Massachusetts, 1988, The MIT Press.

[18] Eric Krotkov and Roberta Klatzky, "Robotic Perception of Material: Experiments with Shape-Invariant Acoustic Measures of Material Type," in *Preprints of the Fourth International Symposium on Experimental Robotics, ISER '95*, Stanford, California, 1995.

[19] R. L. Klatzky, D. K. Pai, and E. P. Krotkov, "Hearing material: Perception of material from contact sounds," *PRESENCE: Teleoperators and Virtual Environments*, vol. 9, no. 4, pp. 399–410, 2000.

[20] Kees van den Doel, D. K. Pai, T. Adam, L. Kortchmar, and K. Pichora-Fuller, "Measurements of Perceptual Quality of Contact Sound Models," in *Proceedings of the International Conference on Auditory Display 2002*, Kyoto, Japan, 2002.

[21] Kees van den Doel and D. K. Pai, "JASS: A Java Audio Synthesis System for Programmers," in *Proceedings of the International Conference on Auditory Display 2001*, Helsinki, Finland, 2001.