VOSIS: a Multi-touch Image Sonification Interface

Ryan McGee
Media Arts and Technology
University of California, Santa Barbara
ryan@mat.ucsb.edu

ABSTRACT

VOSIS is an interactive image sonification interface that creates complex wavetables by raster scanning greyscale image pixel data. Using a multi-touch screen to play image regions of unique frequency content rather than a linear scale of frequencies, it becomes a unique performance tool for experimental and visual music. A number of image filters controlled by multi-touch gestures add variation to the sound palette. On a mobile device, parameters controlled by the accelerometer add another layer expressivity to the resulting audio-visual montages.

Keywords

image sonification, multi-touch, visual music

1. INTRODUCTION

VOSIS began as a custom piece of software necessary to realize a multimedia installation entitled *Voice of Sisyphus*, which explored the sonification of black and white photographs by the artist George Legrady. What began as a simple tool for the audification of image pixel data eventually grew into a parameter-rich, polyphonic sound generator that could be completely controlled via Open Sound Control (OSC). For the installation, an OSC sequencer controlled complex movements of an audio-visual composition. Eventually, the question arose about the potential of performing such a piece live, and it was quickly realized that the use of a computer mouse was not at all conducive to manipulating multiple regions of an image simultaneously. Multi-touch screens provided the ability to move and manipulate the parameters of multiple image regions for a performance scenario, thus creating an instrument. VOSIS currently runs as both an iPad application and a desktop application that can be used with large multi-touch overlays such as those by PQ Labs[1].

1.1 Background

Most experiments examining the relationships between sound and image begin with sounds or music that influence the visuals. Chladni's famous 18th century "sound figure" experiment involves visual patterns generated by playing a violin bow against a plate of glass covered in sand[3]. 20th century visual music artists often worked by tediously synchronizing visuals to preexisting music. Though, in some cases, the sounds and visuals were composed together as in *Tarantella* by Mary Ellen Bute. Today, visual artists often use sound as input to produce audio-reactive visualizations of music in real-time.

Less common are technical methodologies requiring images as input to generate sound. However, in 1929 Fritz Winckel conducted an experiment in which he was able to receive and listen to television signals over a radio[3], thus resulting in an early form of image audification. Rudolph Pfenninger's *Tonende Handschrift* (Sounding Handwriting), Oskar Fischinger's *Ornament Sound Experiments*, and Norman McLaren's *Synchromy* utilized a technique of drawing on film soundtracks by hand to synthesize sounds. VOSIS continues in the tradition of the aforementioned works by using visual information to produce sound.

1.2 Related Work

A vast majority of existing image sonification software uses the so-called "time-frequency" approach [2] in which an image acts as the spectrograph for a sound. These systems include Iannis Xenakis' UPIC and popular commercial software such as MetaSynth and Adobe Audition. Their shared approach considers the entire image much like a musical score where the vertical axis directly corresponds to frequency and the horizontal axis to time. Usually the image is drawn, but some software like Audition allows the use of bitmap images and considers color as the intensity of frequencies on the vertical axis. MetaSynth uses the color of drawn images to represent the stereo position of the sound. In any case, all of the aforementioned software reads images left to right at a rate corresponding to the tempo. Reading an entire image left-to-right as a means to image sonification has been termed as *scanning* by Yeo[5].

However, the approach with VOSIS was to focus on different regions within an image over the course of the performance or composition. Yeo has termed this approach *probing* [5]. Thus, unlike scanning, the horizontal axis of the image is not related to time. The peformer's *probing* of regions over time advances the composition non-linearly. The goal is a more literal translation of images to sound than the typical spectrograph scanning approach. It is felt that, although novel in their own right, spectrograph scanning approaches adhere too closely to a traditional musical score. VOSIS is a departure from the common practice of viewing images as time-frequency planes and provides a technique to listen to variations between different regions of an image. Resulting sounds unfold as one explores areas of an image in a non-linear fashion– first noticing some region, person, or object and then shifting the focus to other objects within the scene.

One convenient constraint with the initial project moti-

[1] `http://multi-touch-screen.com`

NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.

vating VOSIS was that the software did not have to consider color since all source images were greyscale. The possible color-sound relationships with image sonification are numerous and, for now, beyond the scope of this project. While VOSIS has evolved into a more generic tool that can read any image or video, the lack of color consideration remains to serve as a useful limit of scope until current methods are refined and the multiple dimensions of color can be added to the algorithm without causing "parameter overload."

2. IMPLEMENTATION

VOSIS is designed to be used with any image– moving, still, live, or recorded. For still images, the image itself is the "master" image, while for recorded or streaming video, the master image changes at the frame rate of the video or camera. Greyscale pixel values within a region of the master image are read into an array, filtered, drawn as a new image, and read as an audio wavetable. Regions can be selected either by drawing rectangular boxes on the screen or simply by touching an area of an image with a selectable region size. Each region can be thought of as a note with its frequency dependent on the frequency content of the pixels within that region of the image and the rate at which the image pixels are scanned. Chords can be formed by selecting multiple regions at once. There is also a segmentation mode which subdivides large regions and plays them back as a sequence of notes. Consideration was taken for real-time manipulation of region locations and sizes during a performance or installation without introducing unwanted audio artifacts.

2.1 Interface

Users start by selecting an input source, either an image file, video file, or video camera. Regions are added in either "draw" mode or "touch" mode. In draw mode the user can touch to draw outlined rectangular regions of any size. Sounds from regions are looped indefinitely and up to 10 regions can be added in draw mode. Once a region is added its filter parameters (described in detail in section 2.2) can be adjusted via sliders on a GUI panel or mapped to be controlled by a multi-touch gesture or accelerometer direction (when running on an iPad). With multiple regions, touching within a region will make its parameters available for editing on the GUI panel. Thus, it is possible to draw several static regions with various parameters to create a dense sound. In touch mode the user simply touches anywhere on the image to play the sound of an adjustable-sized region centered at the touch location. An ADSR envelope turns touch mode into a keyboard-like instrument. When in touch mode the parameters on the GUI panel will affect every region added via touch mode. So, a typical usage scenario may be to add several regions in draw mode to build up a foundation on which to improvise while in touch mode.

Sounds resulting from regions added to a video or camera feed are inherently dynamic as the content of the region changes at the video frame rate. With both still images and video sources dynamics can be achieved by moving regions to hear the varying amplitude and frequency content within of different areas of the master image. Mapping the filter parameters to multi-touch gestures such as two-finger horizontal and vertical movement and pinch in and out can create a familiar type of sound performance interaction akin to an XY pad or pitch or modulation wheel. Mapping a regions level or scan rate to the iPads accelerometer can result in expressive tremolo and vibrato of the multi-touch "chords." Segmentation mode subdivides regions over an adjustable size threshold into equally sized smaller regions and plays their sounds back in left-to-right, top-down order similar to a step sequencer at an adjustable tempo.

Presentation mode removes the GUI panel and region outlines from sight, making the application suitable for visual music performance. For instance, an iPad can be connected to a projector so that the montage of audio-visual movements can be displayed to an audience without the performers hands covering any visuals. The user is also able to adjust the opacity of the background master image source, making it possible to only display visuals as they are created and played via touch. Thus, it is easy to create live visual music performances by probing and filtering image regions. Figure 1 shows performance mode in comparison to the standard editing mode.

2.2 Synthesis Technique

Figure 2 outlines the image-to-sound synthesis algorithm in VOSIS. As described in section 1.2 VOSIS only deals with greyscale images, and any color or other format images imported to the software will first be converted to 8-bit greyscale. Once a region is selected, the synthesis algorithm begins with a back-and-forth, top-down raster scanning of the greyscale pixel values, which range from 0 to 255 (black to white respectively). Simply scaling these values to obtain a waveform of floating-point audio samples in the -1.0 to 1.0 range results in harsh, noisy sounds without much variation between separate regions in most images. These initial noisy results were not at all surprising given that the greyscale variation of an arbitrary image will contain a dense, broad range of frequencies. For instance, given a picture of a landscape, analyzing variations in each pixel value over a region containing thousands of blades of grass would easily produce a noisy spectrum with no clear partials. Of course, images could be specifically produced to contain particular spectra and result in tonal sounds[6], but the interest of this project is to explore the sounds resulting from different regions of any arbitrary image. When initially experimenting with this form of image sonification one might be tempted to ask "What does a face sound like compared to a window?" However, the ability to determine high-level descriptions of image regions such as a "face" or "window" is a problem of feature recognition in computer vision, and not contained in the scope of this project. Rather, VOSIS examines the objective differences in the pixel data between faces and windows rather than what sounds someone may associate with each of those objects. In many cases the spectral-based filters in VOSIS help to produce less noisy sounds with greater distinguishability between regions.

A selection of frequency domain filters is applied to the audification of pixel data by implementing a short-time Fourier transform (STFT) for each region. The STFT is obtained by computing a fast Fourier transform (FFT) of each region at the graphics' frame rate. Each FFT gives amplitudes and phases for frequencies contained in that region at that time. Manipulation of these amplitudes and phases allows controls the spectrum of the image and, therefore, the resulting sound in real-time. Zeroing the amplitudes of frequencies above or below a cutoff produces a low-pass or high-pass filter respectively, while scrambling the phases of an FFT scrambles the pixels in an image and removes temporal cues from the sound without affecting is spectrum. The key filter was the implementation of a variable amplitude threshold, below which all frequencies are removed, thus leaving only the most prominent partials present to accentuate tonal differences between otherwise similar sounding image regions. Implementing this threshold denoises the resulting sounds, leaving clear tones that change as the region is moved or resized. The pixel data of regions is continuously updated

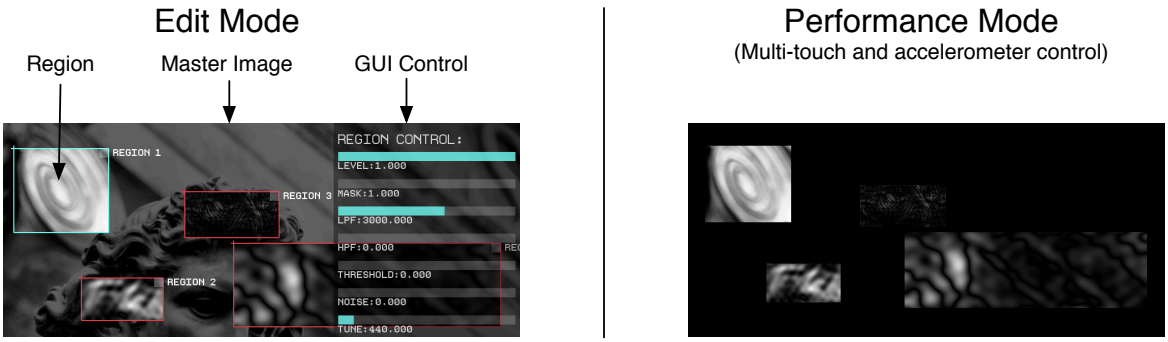

Figure 1: VOSIS Interface: Edit Mode and Performance Mode

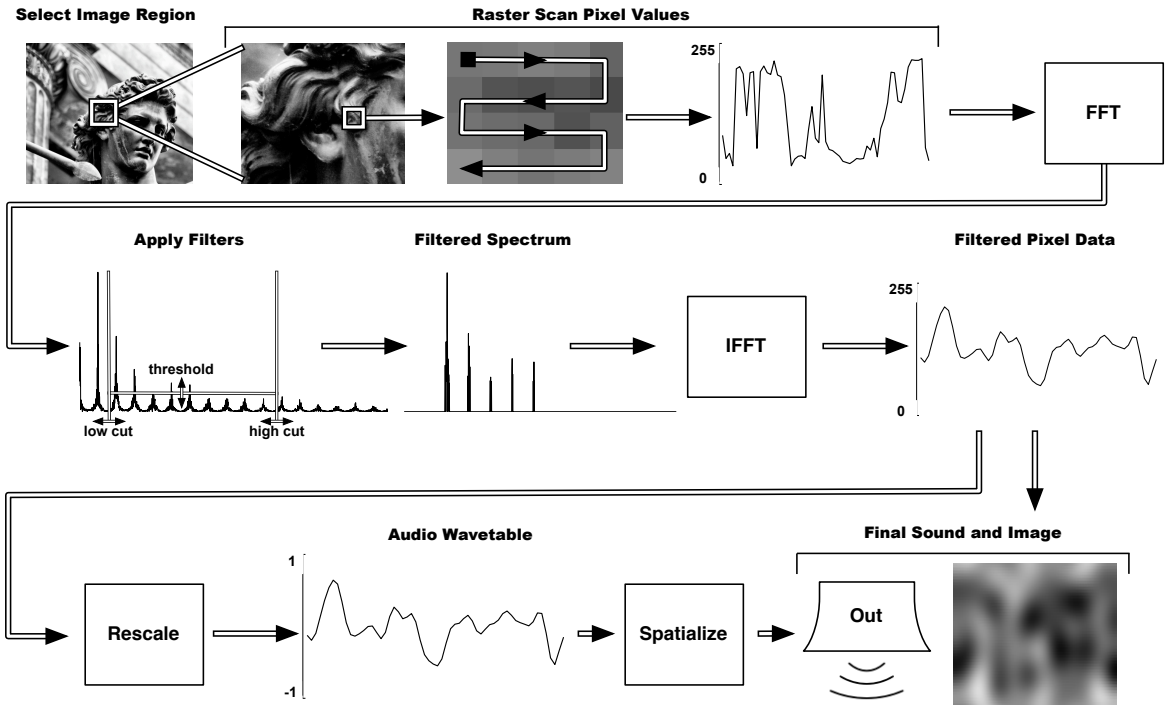

Figure 2: Sound Synthesis Algorithm in VOSIS

to show the effect of the filters so the observer is always seeing and hearing the same data. As the sound becomes clearer from the removal of frequencies, the image becomes blurry. An interesting conclusion from this process is that most perceptually coherent images sound like noise while perceptually clear, tonal sounds result from very abstract or blurry images.

To obtain the final image and sound data after applying filters in the frequency domain an inverse short-time Fourier transform (ISTFT) is computed for each region, which gives the filtered pixel values. These new values are then scaled to the range -1.0 to 1.0 and read as an audio wavetable via scanned synthesis, a technique that can be used to scan arbitrary wavetables of audio data at variable rates using interpolation[4]. A control for the scan rate of these wavetables affects the fundamental pitch of the resulting sounds. However, the perceived pitch also changes as regions are moved and resized, causing new partials appear and disappear from the spectrum.

Before computing the FFT the pixel data can also be scaled to effect the brightness of the resulting image and, therefore, amplitude of the sound. A masking effect can also be applied at this point, which acts as a bit reduction to the image and sound by quantizing amplitude values. Overall, it is important to note that the software only manipulates the image data and not the audio data. Since the audio data is continually produced in the same manner (scanning the IFFT results), changes in the sound are always directly produced from changes in the image. Simply put, using VOSIS one is always seeing and hearing the same data. Figure 3 summarizes the sonic effects of the image filters.

Performance with VOSIS demands rapid movement and resizing of regions which initially caused discontinuities in the wavetables, resulting in an unwanted audible popping noise. To account for the resizing of images, all resulting audio wavetables, originally a length equal to the number of pixels in an image region, are resampled to a fixed size before an interpolated read of the table at the desired frequency. Wavetables are then cross-faded with each other at the audio buffer rate to prevent discontinuities from the dynamically changing wavetables resulting from the movement and resizing of regions. If the regionÕs position and

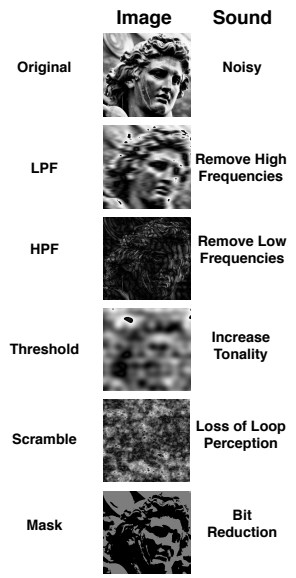

Figure 3: Effects of Image Filters on Sound

size are unchanged, then the wavetable is simply looped. Scrambling the phases of an image region can be used to obtain perceptually continuous sounds rather than loops.

Regions' sounds are spatialized according to their location within the image. If a region is segmented, then the spatialization algorithm updates the position of the sound as each subsection is played. The method of spatialization is similar to that used in vOICe[1], an augmented reality project for the totally blind. Sounds are stereo panned left-to-right according to their region's position in the horizontal image plane. With multiple regions present, the spatialization gives clarity to the mix and provides cues as to the location of sounds within the master image.

3. FUTURE WORK

In continuing work with VOSIS it will be interesting to explore the linked audio-visual effects of typical sound effects elements such as feedback delay lines and arpeggiation. One can imagine more interesting visual music resulting from animated region opacities synchronized with audible delays. Adding an arpeggiator to touch mode will likewise add more sound-synched visual animation. The addition of simple physics for "throwing" regions around the screen via touch may lead to interesting sequencing techniques. Of course, color may eventually be considered to further expand the audio-visual palette.

VOSIS may also have potential as a teaching tool. Undergraduate music and art students in a Processing and Arduino course at the University of California, Santa Barbara expressed a deep curiosity to learn more about Fourier transforms, filtering, image processing, and sound synthesis after playing with VOSIS. The application provides a platform to quickly demonstrate these concepts to students from both visual and music backgrounds.

4. REFERENCES

[1] W. Jones. Sight for Sore Ears. *IEEE Spectrum*, February, 2004.

[2] C. Roads. Graphic Sound Synthesis. *The Computer Music Tutorial*, pages 329–334, 1996.

[3] B. Schneider. On Hearing Eyes and Seeing Ears: A Media Aesthetics of Relationships Between Sound and Image. *See this Sound: Audiovisuology 2*, pages 174–199, 2011.

[4] B. Verplank, M. Mathews, and R. Shaw. Scanned Synthesis. In *International Computer Music Conference*, 2000.

[5] W. S. Yeo and J. Berger. A Framework for Designing Image Sonification Methods. In *Proceedings of International Conference on Auditory Display*, 2005.

[6] W. S. Yeo and J. Berger. Raster Scanning : A New Approach to Image Sonification, Sound Visualization, Sound Analysis And Synthesis. In *Proceedings of the International Computer Music Conference*, 2006.