SAWTOOTH: INTERFACE AS VISUALIZATION

Christopher Burns

University of Wisconsin-Milwaukee
Department of Music
cburns@uwm.edu

ABSTRACT

Sawtooth is an audiovisual improvisation environment in which an information-rich visual interface for the performer simultaneously serves as an aesthetic visual experience for the audience. Gestural input is captured by video camera and translated into audiovisual performance through direct mappings and the initiation of autonomous algorithmic processes. This paper describes *Sawtooth*'s design, with particular attention to the integration of gesture capture, visual feedback to the performer, and audience visual experience.

1. INTRODUCTION

Sawtooth, a new environment for audiovisual performance, addresses the visual aesthetics of live electroacoustic music performance through two distinct means. First, the performer controls the software entirely through physical gestures, typically made with the hands and head, which are captured by a video camera. These gestures invoke the poetics of movement and dance. Second, the animation output of the software is designed to provide both necessary visual feedback to the performer and a compelling visual experience for the audience through the same video channel. *Sawtooth*'s tight integration of gestural interface, visual feedback, and audiovisual performance suggests new potentials for visual interface design, targeted not only at the interface user/performer, but also as an aesthetic experience for audience members. *Sawtooth*'s animation component simultaneously guides the performer's physical gestures, amplifies those gestures for aesthetic effect, and helps to explain the interactive function of those gestures to the audience.

2. RELATED WORK

There are a number of precedents for the concept of the "aesthetic interface." Gaming interfaces share the goal of presenting user feedback at the same time as they provide an immersive experience; however, *Sawtooth* eschews the icons, text, and "heads-up display" metaphors typical of such interfaces [2]. In the domain of computer music, Sergi Jordà's *FMOL* instrument, and its successor project *reacTable*, both use the display of audio waveforms not only as meaningful feedback for the performer, but also as interesting visualizations for the audience [5]. *Sawtooth* differs from these projects in that the visualization is derived from the gesture capture, rather than the audio process. In this regard, *Sawtooth* is more akin to dance performances involving real-time motion capture, though the visualizations generated by such systems are rarely used for performer feedback [1,11]. From a visual art perspective, artists like Myron Krueger, Daniel Rozin, Scott Snibbe, and Camille Utterback have explored similar blurrings of interface and visualization. Golan Levin's "painterly interface" works such as *Loom* are particularly relevant in their concern for the joint specification of sound and image; *Sawtooth* extends this approach by enabling the performer to interact with multiple areas of the canvas simultaneously [7]. As a performance deriving both sound and visuals from hand gestures, Levin and Zachary Lieberman's *The Manual Input Sessions* is also a close relative, though its vignette-based forms are quite different from the improvisational emphasis of *Sawtooth* [8].

3. MOTION CAPTURE

Sawtooth was created to facilitate improvised audiovisual performance, to explore gesture-image-sound couplings, and to investigate the challenges involved in providing a unified visual mode for both the performer's interface feedback and for the audience's visual experience. Input to the system is conducted entirely through video capture of the performer's hand motions and other movements. The system captures video at 64 X 48 pixels of resolution and 30 frames per second (resolution is deliberately scaled down to the working level of detail, to decrease computation load during analysis). Video capture, gesture analysis, and visualization are implemented in Processing; Open Sound Control messages convey gesture data to Pd for audio synthesis rendering [9,10,12].

The primary mode of analysis is motion detection using frame differencing, a technique which is robust over a wide variety of environments and lighting conditions [6]. A variety of secondary analyses are conducted on the motion data, including moving average comparisons of the amount of motion over time, tracking the presence of motion in large groups of adjacent pixels, and tracking the recurrence frequency of motion in each pixel of the frame.

4. THE PERFORMER'S INTERFACE

The choice of video-based motion capture for performer input led naturally (though not inevitably) to the development of a multi-point interface, in which the performer's movement activates different regions of a two-dimensional plane. The presence of motion in different areas of the capture plane is directly translated to the location of visual results on the animation canvas; pitch and timbral features of the audio are also tightly coupled to the x-y locations of motion in the video capture frame.

The first level of interface feedback provided to the performer is a direct display of the motion detection process, helping the performer to perceive the workings of the motion analysis, and to tune the size and speed of her movements accordingly. Each time the frame differencing algorithm determines the presence of motion in a camera pixel, the display draws a white triangle to the screen in the corresponding location of the output video frame (scaled to 1024 X 768 pixels). Each location, or "tile", is flipped left-to-right so that the screen responds to motion as though it were an optical mirror – a more intuitive situation for a performer facing a laptop screen than the native perspective of the camera would offer. A short-duration synthesis event (a sinusoid passed through "soft clip" waveshaping, with variable pregain) is also produced for each detection. Pitch is mapped to the y-axis, with a variety of microtonal scales presented along the x-axis creating a complex two-dimensional pitch "terrain".

The particulars of the waveshaping pregain and the pitch mappings vary over time in response to secondary motion analysis: rapid or large movements decrease the pregain, while infrequent motion increases it (adjusting the brightness of timbre). Significant changes in the overall amount of motion over time (whether increasing or decreasing) produce changes to pitch and scale layouts. These parameter changes are not reflected in the visual display (though the timbral behavior is highly predictable), and must be discovered by triggering audio behaviors.

The second level of performer feedback concerns the tracking of recurrent localized motion. The white triangles reflecting the basic motion tracking gradually fade to black. While they are displayed, their x-y location is "sensitized" for additional motion input. A second motion through the location adds a second white triangle below the first (forming a white square) and activates a timer. A third pass-through ("triple activation") arriving between 800 and 3200 milliseconds after the second creates periodic audio and animation events based on the interonset time between the second and third passes. After the second pass, at the 800 millisecond mark the lower triangle changes color to indicate availability for a third pass-through; the color of the upper triangle gradually changes over the next 2400 milliseconds, so that at the end of the activation period, the square (no longer capable of activation) is a solid nonwhite color. Combined with the multi-point interface, this behavior produces complex and engaging visual and sonic polyrhythmic textures.

Additional color pairings provide information when the performer initiates these autonomous periodic behaviors. The lower triangle fades to black between each periodic "pulse"; the speed of the fade corresponds with the duration of each pulse. Meanwhile, the upper triangle fades to black more slowly, indicating the overall duration of the pulsing behavior. The meanings of the various color relationships between the upper and lower triangles corresponding to a camera pixel are challenging to describe in prose, but easy to grasp with practice; these pairings constitute a key aspect of the visual feedback for the performer. Figure 1 provides an example of the visual appearance of these basic behaviors.

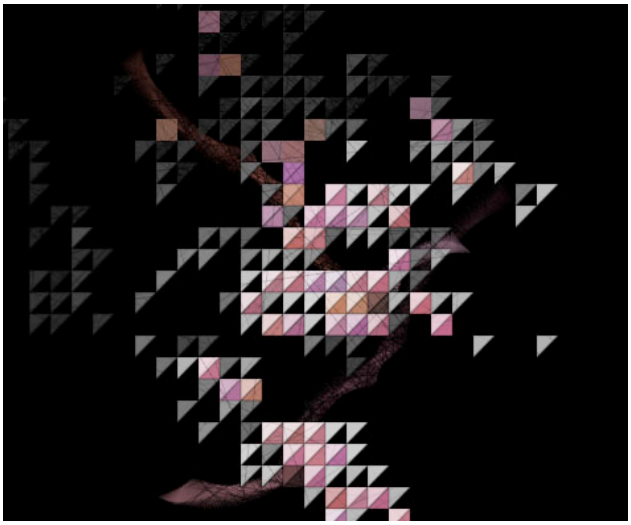

Figure 1. Detail showing the basic gestural modes and visualization of the motion capture.

Identical motion and visual behaviors are also provided at a second level of scale; if ten or more adjacent triangles are activated within a short period of time (implemented with a gradually decaying memory for activations), then large triangles, squares, and periodic events can be activated. These large tiles are arranged in an 8 X 6 grid, and are displayed using a color palette which evolves independently from the small tiles' palette. The three-stage activation process described for the smaller tiles also applies to the larger scale, with the same paired-triangle visualization of the current state and temporal features. When large tiles are activated, the small tiles enclosed by them are rendered temporarily inert and unresponsive, so that the overall activity level of the system need not be driven to high intensity every time a large tile is activated.

Unlike the small tiles, the activation of large tiles is not always immediately reflected in an audio event; the sonic function of these tiles lies in their relationship with the periodic activations of the small tiles, and with one another

(described in greater detail below). Figure 2 provides an example of large and small tiles activated simultaneously.

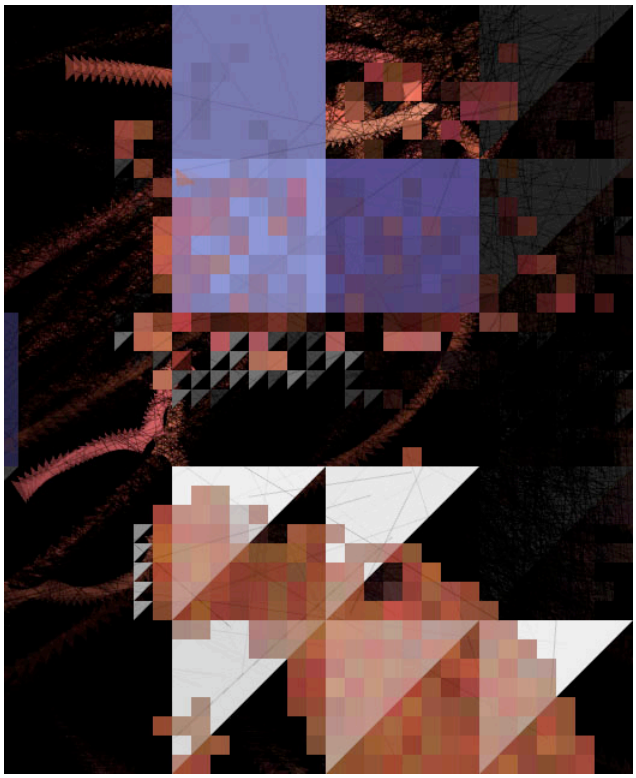

Figure 2. Detail with multiple scales of gesture capture. Note the (gradually decaying) curved traces of moving triangles generated by periodic events.

Additional gestural controls govern the audiovisual relationships between the small-scale periodic events and the large-scale tiles. Each time a small-scale periodic event occurs (after initiation through a triple activation), it launches a moving triangle animation across the visual field. These moving triangles are attracted (via a gravity simulation) to the most recently detected location of motion. The curving traces left by their trajectories are visible in Figure 2.

As a moving triangle crosses the boundaries of activated large-scale tiles, additional synthesis events are created. Complex audiovisual gestures and textures can be created by initiating periodic generation of moving triangles, and then directing the flow of those triangles across active large tiles – an intuitive and viscerally satisfying form of control. Pitch is derived from the initial location of the moving triangle within the two-dimensional pitch terrain established for the small tiles, while durations reflect the speed with which a moving triangle traverses the extents of a particular large tile. Synthesis parameters vary over time in an oscillating manner, if a periodicity has been established for the large tile via multiple pass-throughs.

In this scheme, each large tile offers a unique timbre, through a combination of synthesis and post-processing types. Eight different types of FM synthesis are provided, laid out horizontally across the visual space. The eight varieties differ in the number and relationship of operators, the design of index and amplitude envelopes, and the presence or absence of noise and/or feedback modulation. Each of these synthesis events is passed to a post-processing module, with six types of modules laid out along the vertical axis. Audio processing strategies include vibrato, tremolo, sweeping bandpass filtering, hyperbolic tangent waveshaping, harmonization, and echo. The fixed spatial locations of the synthesis and post-processing types associated with the large tiles facilitate learning and memorization for the performer.

Additionally, the large tile audio modules can be connected by the performer into feedback networks. If multiple adjacent large tiles are driven into periodic behavior (through the triple activation process described above), audio feedback paths are opened between their respective post-processing modules. Input feedback coefficients are at a maximum at the beginning of each period (when the lower triangle of the pair is at maximum brightness) and gradually decay, reaching zero at the end of the period. Invoking different types of post-processing and different degrees of interconnection can produce very different sonic results, from the ringing FM tones produced by the vibrato modules to the very noisy tendencies of the echo post-processors. The feedback connections add another dimension of audio complexity and variety to the system, and provide an compelling sonic analogy for densely active visual states. They also provide an interesting overloading of the triple activation gesture, and further differentiate the audio behaviors of the different gestural and visual scales.

5. THE AUDIENCE EXPERIENCE

The audience sees and hears the same things the performer does. While they may not grasp the informational details of the color pairings, the multiple activation states for each location, or the specific x-y mappings of synthesis types, the connections between gesture, image, and sound that make the visuals function as an interface also serve to clarify the performer's role in the work to the audience. In several cases, animated elements reinforce these mappings. For instance, when the performer directs a flow of moving triangles over a large tile, the large tile flashes in the color of the moving triangles, associating the performance gesture (guiding the moving objects) with the sounding result (new synthesis events associated with the large tile). The more complex and difficult-to-discern relationships between gesture, audio, and visuals facilitate surprise, complexity, and continued engagement with the performance. (The near-obsessive use of triangular forms in the work also tends to emphasize aesthetic unification

through a limited palette of forms, thereby "hiding" the information-bearing aspects of the paired triangles).

There are a number of details of the visualization which do not convey feedback to the performer, but exist only to enhance the aesthetics of the audience experience. First, the color palettes of the small and large tiles evolve independently, autonomously, and on separate time scales. (Color relationships bear information about gesture and state, but the specific color choices do not). Second, the background of the image is not refreshed with each new frame; instead, a series of black "film scratch" lines are drawn over the previous frame, gradually degrading visual elements which aren't redrawn with each new frame (most notably the moving triangles associated with periodic events). Several parameters of this "film scratch" process are time-varying, creating different effects and senses of "memory" and "decay" over the course of a performance.

Similarly, there are aspects of the audio environment which evolve without any visual display to the performer. The pitch terrain of small tile activations, and a variety of parameters for large tile synthesis events, depend upon performer discovery and improvisational adaptation to the current range of possibilities. These time-varying mappings are not only aimed at the audience (they have an enormous effect on the performer's choices), but they do function in parallel to the autonomous visual evolutions.

6. ASSESSMENT

Sergi Jordà's criteria for the evaluation of digital musical instruments – learnability, efficiency, diversity, and expressivity – are all appropriate for the assessment of *Sawtooth* [3,4]. With regards to diversity, *Sawtooth* performance is limited to its own genre – the software is an "interactive composition" or "improvisation environment" rather than an instrument capable of performance in many styles. However, *Sawtooth* does offer both performance diversity (performances can be made substantially different from one another, especially in terms of sonics and formal design) and performance nuance (expressive differentiation between trained and untrained performers is made through the effective use of physical gesture, subtle deployment of overloaded gestures and the multi-point capability of the interface, and the creation of convincing large-scale form).

Sawtooth performs well in learnability and efficiency: the same direct mappings that help the audience to perceive the workings of the software invite the performer up the learning curve, and the visual feedback of the interface is easy to grasp and deploy in the service of more complex audiovisual gestures. While it sounds hopelessly overwhelming to say that *Sawtooth* can display up to 3120 unique states across its animation field simultaneously, in practice the informative aspects of the animation are quickly internalized. When set up in an installation-style presentation, visitors seem to enjoy performing with the work. They discover and learn its various behaviors and capabilities without instruction, and they devote significant time to the experience – all promising indications.

In terms of expressivity, the software engages repeated rehearsal and performance, and provides an effective balance between reproducible and nonlinear behaviors for improvisation. The integration of gesture, display, and performance in *Sawtooth* leads to new artistic potentials, and suggests that information display and aesthetic experience can be mutually enriching concepts in the design and implementation of audiovisual performance.

7. REFERENCES

[1] Camurri, A., B. Mazzarino, and G. Volpe, "Expressive Gestural Control of Sound and Visual Output in Multimodal Interactive Systems", in *Proceedings of the International Conference Sound and Music Computing*, Paris, France, 2004.

[2] Dyck, J. et. al. "Learning from games: HCI design innovations in entertainment software," in *Proceedings of Graphics Interface*, Halifax, Nova Scotia, 2003.

[3] Jordà, S. "Digital Instruments and Players: Part I – Efficiency and Apprenticeship", in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Hamamatsu, Japan, 2004.

[4] Jordà, S. "Digital Instruments and Players: Part II – Diversity, Freedom, and Control", in *Proceedings of the International Computer Music Conference*, Miami, USA, 2004.

[5] Jordà, S. "Interactive Music Systems for Everyone: Exploring Visual Feedback as a Way for Creating More Intuitive, Efficient, and Learnable Instruments", in *Proceedings of the Stockholm Musical Acoustics Conference*, Sweden, 2003.

[6] Levin, G. "Computer vision for artists and designers: pedagogic tools and techniques for novice programmers", in *AI & Society* 20/4, 2006, 462-482.

[7] Levin, G. "Painterly Interfaces for Audiovisual Performance." M.S. Thesis, MIT Media Lab, 2000.

[8] Levin, G. and Lieberman, Z. "Sounds from Shapes: Audiovisual Performance with Hand Silhouette Contours in *The Manual Input Sessions*", in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Vancouver, Canada, 2005.

[9] Puckette, M. "Pure Data", in *Proceedings of the International Computer Music Conference*, Hong Kong, China, 1996.

[10] Reas, C. and B. Fry. *Processing*. Cambridge: MIT Press, 2007.

[11] Vanier, L., H. Kaczmarski, and L. Chong. "Forming the dots: live optical motion capture dance", in *ACM SIGGRAPH 2003 Sketches & Applications*.

[12] Wright, M. "Open Sound Control: An enabling technology for musical networking", in *Organised Sound* 10(3), 2005, 193-200.