Sounds from Shapes: Audiovisual Performance with Hand Silhouette Contours in *The Manual Input Sessions*

Golan Levin
Carnegie Mellon University
College of Fine Arts, CFA-300
5000 Forbes Avenue
Pittsburgh, PA, 15213 USA
+1.412.268.2000
golan@andrew.cmu.edu

Zachary Lieberman
Parsons School of Design
Design and Technology Dept.
2 W. 13th Street, 10th Floor
New York City, NY, 10011 USA
+1.212.229.8908
zlieb@parsons.edu

ABSTRACT
We report on *The Manual Input Sessions,* a series of audiovisual vignettes which probe the expressive possibilities of free-form hand gestures. Performed on a hybrid projection system which combines a traditional analog overhead projector and a digital PC video projector, our vision-based software instruments generate dynamic sounds and graphics solely in response to the forms and movements of the silhouette contours of the user's hands. Interactions and audiovisual mappings which make use of both positive (exterior) and negative (interior) contours are discussed.

Keywords
Audiovisual performance, hand silhouettes, computer vision, contour analysis, sound-image relationships, augmented reality.

1. INTRODUCTION
It is easy to understand how the *hand*, as one of the most highly articulated, neurologically sensitive, and proprioceptively adept parts of the body, has come to have such a primary role in both musical communication – as our means for performing nearly all musical instruments – and in live visual communication, through expressive forms such as shadow play and sign language. Hands are quite simply very well-adapted to communicative expression in both the audible and visible domains.

It is impossible to estimate when humans first entertained each other with plays of hand shadows, or first used their hands to bring forth sounds from a musical instrument. Evidence certainly points to the possibility that both activities have been a part of human culture for many thousands of years. But to the best of our knowledge, we are unaware of any traditions or technologies in which the hands are used to *simultaneously* perform both visual shadow-play and instrumental musical sound.

A system for hand-driven audiovisual performance is the goal of the research we present here. More precisely, we report on the design and development of easily learnable, richly expressive software mechanisms by which people can create and perform both animated imagery and sound, simultaneously, in real-time, using only their hands in an unencumbered manner.

Our criteria for success are similar to those we used to develop our previous mouse-driven audiovisual performance systems, the *Audiovisual Environment Suite* [4], and our recent voice-driven audiovisual performance systems, *RE:MARK* and *Messa di Voce* [5]. These criteria include such seemingly contradictory goals as:

- **Simplicity/Difficulty**: the system's basic principles of operation are easy to deduce and self-revealing; at the same time, sophisticated expressions are possible, and true mastery requires the investment of practice.
- **Repeatability/Inexhaustibility**: the system responds consistently to consistent input; and yet, the system never responds exactly the same way twice, because it is sensitive to miniscule differences in user performance.
- **Create, Manipulate, Destroy**: the system provides an audiovisual material for which all three actions are possible.
- **Audiovisual Commensurability**: the system's sonic and visual dimensions are equally malleable.

In developing systems for hand-driven audiovisual performance, we have decided to use computer vision techniques specialized for the detection, identification and analysis of closed silhouetted contours [2]. Such techniques allow for the unfettered extraction of gesturally significant data about the user's hand postures and movements. As we shall see, our instruments use these data to govern the synthesis of both graphics and sound.

2. BACKGROUND
In this section we consider prior research (1) in which vision-tracked hand contours are used as a primary interface for interactive visual play; (2) in which body-driven shape contours, obtained from vision-based analysis of video, have been used to govern the real-time control of expressive musical parameters, and (3) in which sensor-tracked hands are used to control virtual audiovisual objects.

2.1 Myron Krueger's *VIDEOPLACE*

Myron Krueger's influential *VIDEOPLACE* system, created more than thirty years ago, was one of the first interactive artworks to make use of computer vision as a means for capturing and incorporating the gestural expressions of its users. Of particular relevance to our research is the fact that many of Krueger's playful interactions made extensive use of the silhouettes of participants' hands. By detecting and tracking the tips of participants' fingers, for example, Krueger enabled participants to create virtual synthetic drawings in mid-air; to adjust the shape of a virtual Bezier curve; and to experience the fiction of picking up and holding the entire body of a networked cohort in one's hands.

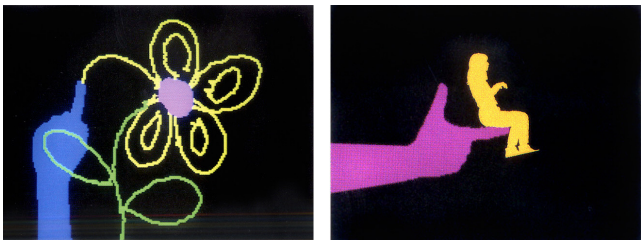

Figure 1. Myron Kruger's *VIDEOPLACE* structured many playful interactions around the use of participant hand silhouettes. Images reproduced from [3].

Clearly, Krueger had by 1975 given a great deal of consideration to the ways in which the computational augmentation of hand silhouettes could be used to prompt both narrative and abstract forms of creative visual play. Information about the extent to which Krueger employed participant contours to govern *sound*, however, is scanty. Although he reports having linked the stereo placement of synthetic sounds to a participant's horizontal position in his early *GLOWFLOW* module [3] (a mapping we also adopt in our current research), little else is known. From his brief mentions of sound in his book *Artificial Reality II*, which stands as the primary document of the *VIDEOPLACE* project, it seems safe to say that the musically instrumental quality of his systems was not a primary focus of his research.

2.2 Lyons et al.'s *Mouthesizer*

In their *Mouthesizer* (2001), Michael Lyons and his colleagues use statistical measurements computed from the visually-tracked contour of a performer's mouth in order to modulate musical parameters of real-time audio synthesizers and filters [6]. In their design, the *Mouthesizer* performer wears a miniature head-mounted camera directed at his mouth. The cavity of the performer's mouth is detected by intensity and color thresholding; the largest detected pixel-blob is then subjected to various morphological analyses in order to distill a small number of highly descriptive shape metrics. In one demonstration, these parameters (which describe e.g. the width, height, and compactness of the mouth aperture) are used to control sound properties such as the cut-off frequency of a resonant low-pass filter, or the distortion level (non-linearity) of an audio amplifier. These dynamic filters are in turn used to modify the sound of a live guitar signal performed by the same person, or a track in a pre-sequenced techno composition.

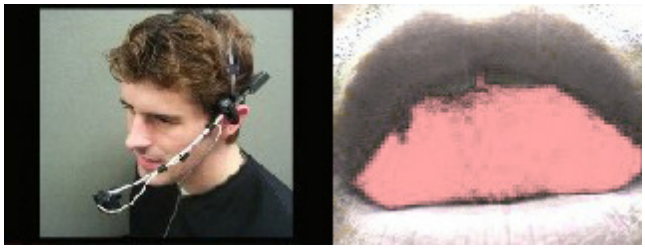

Figure 2. The *Mouthesizer* [6] uses the shape of the mouth to control real-time audio. The pink color indicates the area segmented by its vision system. Images reproduced from [6].

The *Mouthesizer* is relevant to our present research because it demonstrates the technical and instrumental feasibility of visually tracking a body-based contour, and offers mappings according to which such a contour can serve as an intuitive and expressive handle into a musical experience. From the *Mouthesizer*, we have borrowed the idea of tracking an *interior contour* (also called a hole or negative shape), although we direct our attention to those interior contours which can arise between the fingers of the hand.

A significant difference between the *Mouthesizer* and our hand-driven instruments is that Lyons et al. appear only to use the mouth contour to *modulate* or filter musical sound, rather than to *cause* or create it. Consequently, the *Mouthesizer* can only function when it is used in tandem with another musical instrument (such as a guitar) or a pre-recorded musical passage. A second important difference is the extent to which the *Mouthesizer* is intended as a *visual* performance instrument, e.g. for live cinema. Although the *Mouthesizer* readily provides an interesting view of the performer's mouth, its current implementation (as described in [6]) does not allow for the creation or manipulation of graphical phenomena apart from the live view of this contour. Our instruments, by contrast, work independently from other sound sources and allow both sounds and images to be created and manipulated together.

2.3 Mulder et al.'s *Sound Sculpting* Systems

In their *Sound Sculpting* systems (1998-1999), Axel Mulder et al. use hand posture information (captured in real-time by Cyberglove dataglove hardware and Polhemus 6-DOF position sensors) to adjust the shape and location of a virtual 3D object [8]. Various physical properties of this virtual object (such as its size, curvature, and amount of torsional twist) are then used to govern continuous sound effects such as flange strength, chorus depth, FM distortion and vibrato. In addition to the resulting sound, the virtual object (a flexible sheet or balloon) is also visualized on a nearby computer monitor, thus constituting a complete audiovisual display.

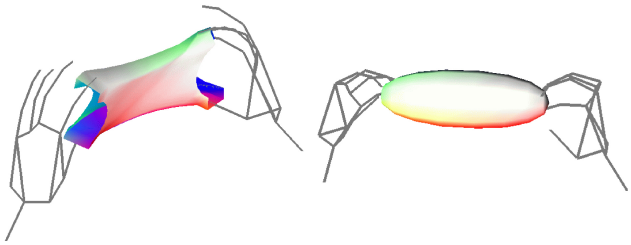

Figure 3. Mulder et al.'s *Sonic Sculpture* systems, showing wireframe hands shaping sound-objects. Images are from [8].

Mulder et al.'s work is a significant precursor to ours because it demonstrates that the hands can be used to manipulate both the audible and visible aspects of virtual objects, simultaneously and in real-time. Superficially, it may seem that the work we present differs only in our selection of hand-tracking technology (we use machine vision instead of the more encumbering datagloves) and in our visual aesthetics (we use 2D virtual objects instead of 3D ones). Nevertheless, there are more significant differences as well.

In Mulder's system, the pitch and duration of all sounds are fixed in a preprogrammed MIDI sequence, while hand postures are used to modulate audio effects applied to this material. In other words, as with the *Mouthesizer*, Mulder's interaction methods pre-suppose the existence of a virtual sound-producing object in the first place. By contrast, our systems permit a user to instrumentally *create* virtual sound-objects, as well as to modify them thereafter.

Because all hand measurements in Mulder's system are mapped to continuous position variables, no actions or states remain available for symbolic or discrete forms of audiovisual control. Thus, for example, despite the considerable expense and sophistication of their glove hardware, Mulder et al. are compelled (somewhat perversely) to rely on a footswitch to enable and disable "holding" of the virtual object. We contend that much more articulate forms of discrete control are essential for the initiation and termination of non-canned musical events. To this end, we employ silhouettes as our primary interface because they support both continuous and discrete logics for interaction: on the one hand, contour shapes can be continually modified; on the other, interior contours can be effortlessly created or destroyed.

3. PERFORMANCE AND INSTRUMENTS

3.1 *The Manual Input Sessions* Performance

The Manual Input Sessions is an audiovisual performance intended to probe the expressive possibilities of unencumbered hand gestures. The concert consists of a suite of custom vision-based software systems, described below, which are performed on a combination of a traditional analog overhead projector, and a modern DLP video projector. In our setup, the analog and digital projectors are registered and aligned such that their projections overlap, resulting in an unusual quality of hybridized, dynamic light. Because the rectangular boundary of the video projection is invisible, the concert presents the apparent fiction that the entire display is produced by a "magical overhead projector."

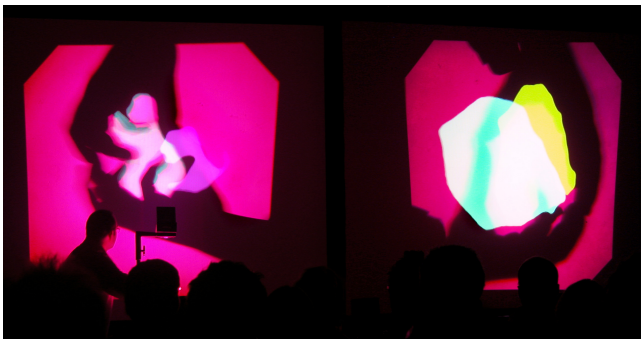

Figure 4. The authors performing *The Manual Input Sessions* at Ars Electronica Festival, Linz, Austria, September 2004.

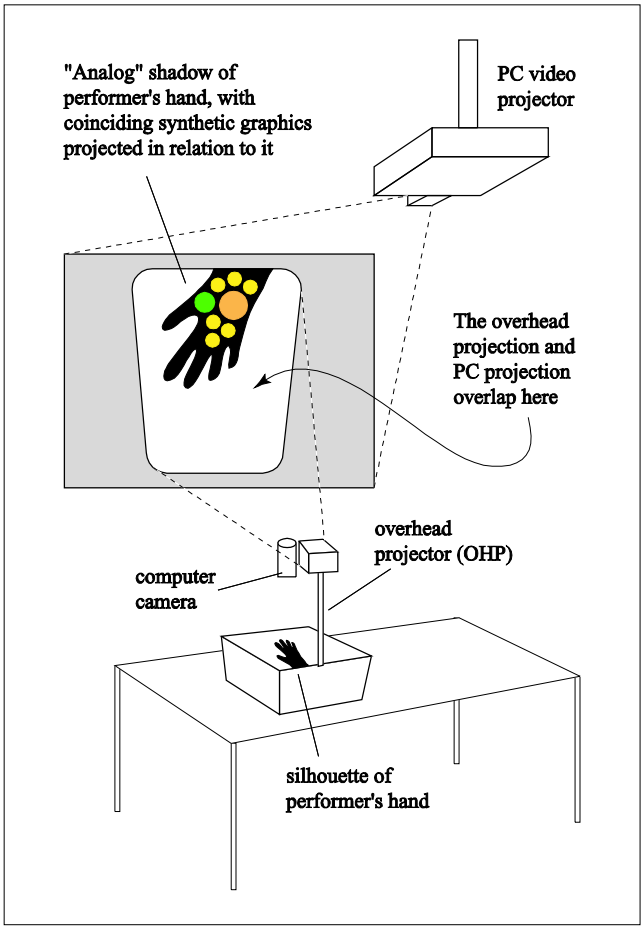

Figure 5. Hardware setup of *The Manual Input Sessions.* The analog and video projections are aligned to coincide exactly.

During the performance, a computer vision system analyses the silhouettes of the performer's hands as they move across the glass platen of the overhead projector. The contours of these silhouettes are then analyzed by our custom instruments. In response, our software generates synthetic graphics and sounds that are tightly coupled to the forms and movements of the performer's hand movements and postures. These synthetic visual responses are co-projected over, into and around the overhead's analog shadows – with which they have been carefully aligned. The result could best be described as a form of "augmented reality shadow play."

This paper discusses three of the software instruments used in our performance: *NegDrop*, *InnerStamp*, and *Rotuni*. Each of these modules creates audiovisual responses to the silhouette contours of the performer's hands. In fact, our full instrumentation includes several other systems not discussed here, such as an instrument which allows performers to influence a granular synthesizer by drawing a force field (with their fingertips). Another sub-system (which makes use of object recognition techniques) allows performers to switch between instruments by placing a cardboard symbol on the platen of the OHP.

To create legible visual contrast between the analog shadows and illuminated digital graphics, we place a sheet of colored theatric gel onto the glass platen of the OHP. This appears as a gray or magenta background in Figures 4, 6, 8, 10, and 12.

3.2 The "NegDrop" Instrument

In our *NegDrop* performance module, closed interior contours (i.e. holes or negative spaces) in the performer's hands are detected by the computer vision system, and used as visual representations of virtual sound-producing objects. (Such interior contours can be made, for example, by enclosing an empty region between one's thumb and forefinger, as with the "OK" hand sign.) When the performer breaks the contour of the hole by separating his fingers, the shape is released from his hand and falls downward as if pulled by gravity.

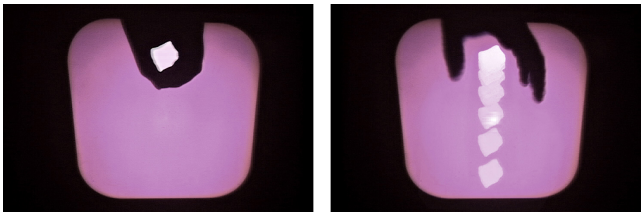

Figure 6. In the *NegDrop* instrument, interior contours become "droppable" virtual objects which trigger sounds when they collide with the boundaries of the projection. [The right-hand photograph is a time-lapse composite.]

When the virtual shape collides with the boundaries of the projection area, it bounces rigidly off the boundary and triggers the production of a MIDI sound whose properties are closely coupled to certain visual aspects of the dropped shape. (The audiovisual mappings in *NegDrop* are given in Table 1.) With each bounce, the dropped object voices its sound and loses a percentage of its kinetic energy to simulated friction; after a while, the object lacks sufficient energy to continue bouncing and is made to fade away. In our current implementation, virtual objects dropped from the top of the projection bounce for approximately five seconds.

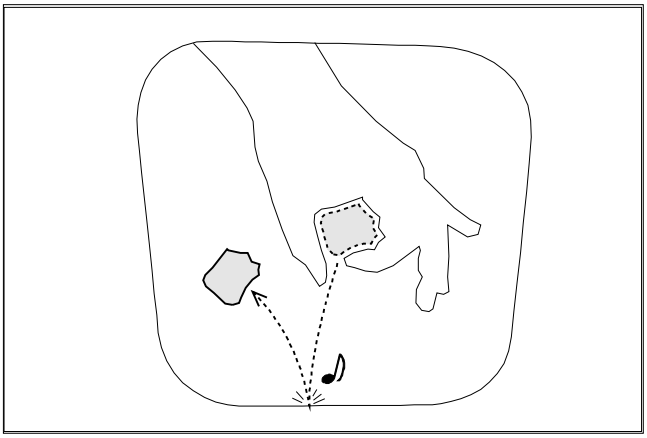

Figure 7. Dropped objects inherit their initial lateral velocity from the horizontal movement of the hand that released them. The horizontal position of the virtual object governs the stereo position of the sounds it produces.

Although the performer can quickly deposit a large number of bouncing virtual shapes ("Neggs"), such that many shapes co-exist in the projection simultaneously, the implementation of inter-shape collisions is currently disabled, as the sounds caused by secondary collisions between Neggs were judged to be too chaotic.

Table 1. Audiovisual Mappings in *NegDrop*.

Contour Properties	Sound Properties
contour area	pitch (large = low)
collision energy	volume
horizontal position	stereo pan location
compactness / pointiness	timbral brightness

Instrumentally speaking, it is somewhat difficult to predict the precise pitch which a dropped Negg will produce. Small variations in shape area, owing to such factors as the variability in the distance from the performer's hand to the glass platen of the OHP, can lead to pitch variations of one or two semitones. The *NegDrop* instrument is consequently a poor choice for the performance of explicitly melodic musical material. At the same time, it is quite easy to predict the *general* pitch range in which a Negg will sound. *NegDrop* additionally affords very precise control of note attack timing, as this can be directly regulated by the distance from the performer's hand to the virtual floor. As a result, *NegDrop* is a good instrument for performing textures of note-clusters and some varieties of pitched rhythmic percussion.

Our current implementation of *NegDrop* uses MIDI as an expedient means of triggering real-time sound events. Owing to *NegDrop*'s use of simulated physics, however, this instrument is a good candidate for the use of physical modeling-based synthesis techniques such as those described by O'Brien et al. in [9]. In such a design, which we intend to pursue in a future version of the *Manual Input Sessions* project, synthetic sounds would be computed by modeling our silhouette-derived virtual objects as elastic masses with shape-specific modes of natural vibration.

3.3 The "InnerStamp" Instrument

Like *NegDrop*, the *InnerStamp* performance module also uses negative contours inside the performer's hands to generate sound. Unlike *NegDrop*, however, *InnerStamp* presents an interaction for the synthesis of continuous drones, rather than the triggering of discrete notes.

When the performer of *InnerStamp* creates a closed negative shape within the silhouette of her hands, this interior contour is highlighted, and a pitched drone is heard. As long as the performer does not rupture the shape's contour, the sound of this drone can be continuously modified by changing various visual properties of the contour. Flattening the contour into a long, thin shape, for example, brightens the timbre of its drone. Changing the perimeter of the shape from large to small causes its drone to rise in pitch.

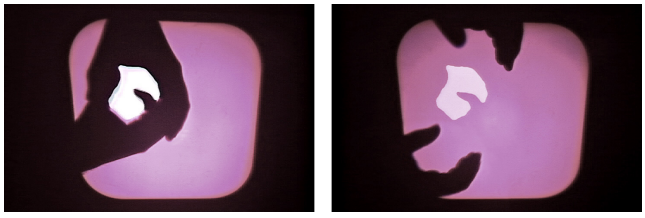

Figure 8. In the *InnerStamp* instrument, interior contours persist after they are created.

InnerStamp uses a hybrid granular/FM synthesizer implemented using the real-time audio affordances of Ross Bencina's *PortAudio* library and Stephen Pope's *CSL* toolkits [1],[11]. *InnerStamp* consequently offers extremely precise control of pitch and timbre. Further details about its mappings can be found in Table 2, below.

Table 2. Audiovisual Mappings in *InnerStamp*.

Contour Properties	Sound Properties
contour perimeter	pitch (large = low)
horizontal position	stereo pan location
time since hands departed	volume decay
perimeter-to-area ratio (i.e. non-compactness)	FM modulation index (i.e. timbral brightness)

A unique aspect of the *InnerStamp* instrument is that, during the time that the user is still holding the negative shape "inside" her hands, the shape records all of the transformations that are happening to it. These transformations include any and all of the user's real-time manipulations of the contour's size, position, or boundary shape. After the user "releases" the shape (by opening up her hands), the shape remains in the projection – and *plays back* the recorded manipulations which happened to it earlier. These transformations replay endlessly, looping back-and-forth, until the user removes her hands from the projection, at which point the contour's sound and image gradually fades away.

While an animating shape replays its morphological history, it also replays its sonic history. Thus a shape which was created to animate from large to small (and hence glide from a low drone to a high-pitched one) will replay this sound-passage while it also loops visually from large to small and back again.

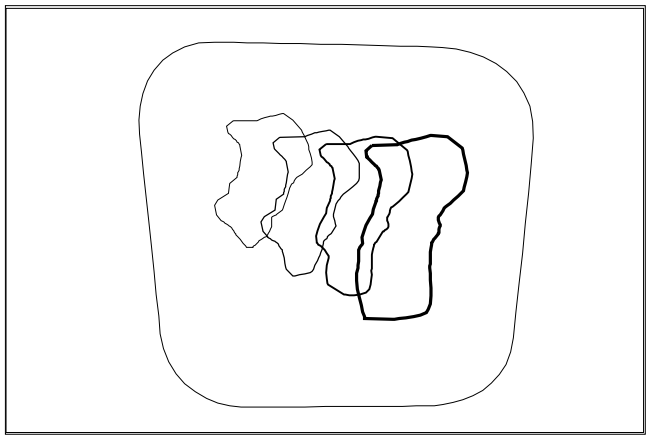

Figure 9. Interior contours deposited into the *InnerStamp* projection replay their individual histories of movement.

The *InnerStamp* instrument permits up to three animating shape-stamps to be deposited into the projection at any one time. (Using more than three simultaneously was judged to be too chaotic.) Each newly-introduced recording replaces the oldest active stamp.

3.4 The "Rotuni" Instrument

The *Rotuni* instrument develops rhythmic melodic ostinatos from the positive contours of the performer's hands, or any other opaque objects which are placed onto the glass platen of the system's overhead projector. Unlike *NegDrop* or *InnerStamp*, it is not necessary for the performer of *Rotuni* to create an interior (negative) contour in order for the system to produce sound.

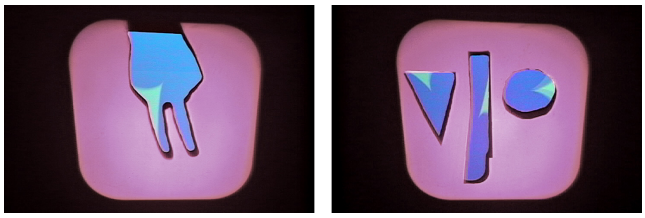

Figure 10. The *Rotuni* instrument generates a rhythmic melody for each positive silhouette contour it identifies.

Users play *Rotuni* by placing their hands or other objects on the glass surface of the overhead projector. The outline contours of the individual objects are individually segmented and tracked by the computer. These silhouettes are then digitally re-projected onto the projection screen, but with the significant addition of a virtual "clock arm" similar to an old-fashioned radar display. This arm extends from the centroid of each silhouette to its edge, and rotates in discrete rhythmic time steps according to a pre-set tempo.

As the clock arm sweeps around the contour of the silhouette, a MIDI note is triggered whose pitch is proportional to the length of the clock arm at that time-step. Thus, for example, circular shapes yield drone-like pulses, while shapes with odd protuberances (like fingers) create high notes when the clock arm sweeps past a finger (and lower notes otherwise). The *Rotuni* is polyphonic, since each silhouette yields its own melody.

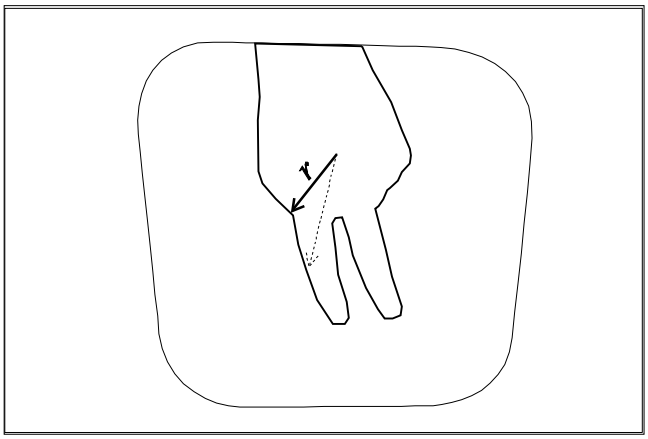

Figure 11. The pitch produced on a given beat is proportional to the length of the shape's rotating radial arm.

Each silhouette, moreover, yields a melody which is unique to its form. Cardboard cutout shapes can be designed, therefore, which yield predictable melodies when placed into the system. In *The Manual Input Sessions* performance, we employ a combination of malleable silhouettes from our hands, fixed cutout cardboard shapes, and everyday objects (such as coins, scissors, keys, and PC mice) when playing the *Rotuni* instrument.

Table 3. Audiovisual Mappings in *Rotuni*.

Contour Properties	Sound Properties
length of sweeping radial arm	pitch (short=low)
horizontal position	stereo pan location
contour ID number	MIDI timbre selection

Rotuni offers an intuitive interface for controlling melodic material in a rhythmic context. It is even possible to perform musical rests in *Rotuni*'s otherwise periodic beat, by creating C-shaped silhouettes whose centroids lie outside the shape's boundary. Regrettably, our current implementation of this instrument does not provide any other interface mechanism for modulating its volume dynamics, or regulating its basic tempo. Although there are obvious non-intrinsic solutions to these issues (e.g. volume pedals and/or keyboard buttons), this is an area of further research for us.

4. CONCLUSIONS

We present several instruments that use the interior and exterior contours of hand silhouettes, as detected and analyzed by a computer vision system, to create and manipulate sound and animated imagery simultaneously. Recognizing Lev Manovich's definition of augmented reality – as an "overlaying of dynamic and context-specific information over the visual field of a user" [7] – we conclude that our instruments, which merge real-time sound with virtual synthetic graphics and organic analog shadows, enable a new form of live audiovisual cinema to be performed in the hybrid locale of an augmented reality.

5. ACKNOWLEDGMENTS

A version of the *Rotuni* instrument was originally created in 1997 at Interval Research Corporation with the collaboration of Scott Snibbe, Marcos Vescovi and Philippe Piernot [10]. Further development of our instruments and performance was made possible through support from the 2004 Whitney Biennial, The Kitchen, the 2004 Ars Electronica Festival, and RomaEuropa Festival 2004. We are indebted to Gregory Shakar, Andrea Boykowycz and Nurit Bar-shai for their invaluable support and assistance with the performances of this project.

6. REFERENCES

[1] Bencina, Ross. *PortAudio* sound synthesis library. http://www.portaudio.com/.

[2] Da Fontoura Costa, L. et al. *Shape Analysis and Classification: Theory and Practice.* CRC Press, 2000.

[3] Krueger, M. *Artificial Reality II.* Addison-Wesley, 1991.

[4] Levin, G. "Painterly Interfaces for Audiovisual Performance." M.S. Thesis, MIT Media Laboratory, August 2000. http://acg.media.mit.edu/people/golan/thesis/.

[5] Levin, G. and Lieberman, Z. "In-Situ Speech Visualization in Real-Time Interactive Installation and Performance." *Proc. 3rd International Symposium on Non-Photorealistic Animation and Rendering*, Annecy, France, 2004.

[6] Lyons, M., Haehnel, M., and Tetsutani, N. "The Mouthesizer: A Facial Gesture Musical Interface." *Conference Abstracts, Siggraph 2001*, Los Angeles, p. 230.

[7] Manovich, Lev. *The Language of New Media.* MIT Press, 2001.

[8] Axel G.E. Mulder, S. Sidney Fels and Kenji Mase. "Design of Virtual 3D Instruments for Musical Interaction." *Proceedings of Graphics Interface '99*, (Kingston, ON, Canada, 2-4 June 1999, S. Mackenzie and J. Stewart (eds.)) pp. 76-83, Toronto, ON, Canada: University of Toronto.

[9] O'Brien, J., Cook, P., and Essl, G. "Synthesizing Sounds from Physically Based Motion." *The proceedings of ACM Siggraph 2001*, Los Angeles, California, pp. 529-536.

[10] Piernot, P., Vescovi, M., Cohen, J., Levin, G., et al. "Video camera based computer input system with interchangeable physical interface" *(A modular tabletop surface for use with computer-vision-based children's games).* US Pats. 5953686 and 6047249. Filed 7 July 1996, issued 4 April 2000.

[11] Pope, Stephen et. al. *CREATE Signal Library (CSL).* http://www.create.ucsb.edu/mailman/listinfo/csl.

Figure 12. Visual summary of the hybrid analog/digital light projection technique used in *The Manual Input Sessions* instruments. Left to right: (1) Live source imagery of the performer's hand silhouettes is obtained from the overhead projector; (2) Hand silhouettes are analyzed by a computer vision sub-system, and computer graphics (typically two-dimensional lines and polygons) are generated in response; (3) The synthetic graphics are warped by an affine transform in order to accommodate any necessary perspective corrections, and then projected so as to coincide with the light projection emitted by the overhead projector.