WaveForm: Remote Video Blending for VJs Using In-Air Multitouch Gestures

Amartya Banerjee
Human Media Lab
Queen's University
Kingston, ON K7L 3N6 Canada
banerjee@cs.queensu.ca

Audrey Girouard
Human Media Lab
Queen's University
Kingston, ON K7L 3N6 Canada
agirouard@gmail.com

Jesse Burstyn
Human Media Lab
Queen's University
Kingston, ON K7L 3N6 Canada
jesse@cs.queensu.ca

Roel Vertegaal
Human Media Lab
Queen's University
Kingston, ON K7L 3N6 Canada
roel@cs.queensu.ca

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
We present WaveForm, a system that enables a Video Jockey (VJ) to directly manipulate video content on a large display on a stage, from a distance. WaveForm implements an in-air multitouch gesture set to layer, blend, scale, rotate, and position video content on the large display. We believe this leads to a more immersive experience for the VJ user, as well as for the audience witnessing the VJ's performance during a live event.

Keywords
Video blending, remote interactions, multitouch, large displays, VJ, video jockey

ACM Classification
H5.2 [Information interfaces and presentation]: User Interfaces: Input Devices and Strategies, Interaction Styles.

General Terms
Design, Human Factors

Introduction
Video Jockeys are artists who create a real-time video performance by manipulating and mixing video imagery. Typically, VJing takes place during an event

Figure 1. VJ Event, Cologne 2005 (Collage [7])

like a music festival, in a nightclub, at a concert etc., with the video performance projected on one or more large screens around the event space (Figure 1). VJs improvise over music tracks by manipulating the video content as well as their properties, such as transparency, orientation or theme.

One of the most important characteristics of VJing is the capacity to have live control over media. VJs create a real-time mix using video content that is pulled from a media library that resides on a laptop hard drive, as well as computer visualizations that are generated on-the-fly. VJs typically rely on a workspace with multiple laptops and/or video mixers set on a table. However, with this setup, there is a disconnect between the instrument, the VJ performance and the visual output. For example, VJs cannot typically see the effect of specific videos on the large screen. In addition, the audience experience of the VJ's creative expression is limited to that of watching her press buttons on her laptop.

To address these issues, we investigated ways in which videos can be directly manipulated on the large screen. To avoid VJs from blocking the screen, we looked at solutions that allowed for remote gestural control of the visual content. In this paper, we discuss WaveForm, a computer vision system that uses in-air multi-touch gestures to enable video manipulation from a distance. Our system supports video translation, resize, scale, blend, as well as layer operations through specific in-air multi-touch gestures. To allow the large display to only be used for rendering the result, the WaveForm VJ uses a tablet computer as a private media palette. VJs select and cue videos and music tracks on this palette through multitouch gestures, which are then transferred to and manipulated on the audience display via in-air gestures.

Related Work

Our work draws from both the culture of VJing and research in distance-based multi-touch systems.

Interviews conducted by Engström et al. [3] provide insight into the approaches VJs take when producing a set. Mashup compositions comprise of continually adding layers to the performance and rapidly playing with the blend and effects. Uniform compositions maintain a theme, gradually perfecting the blend while evolving the visuals over time. They note that currently the interaction between the VJ and the audience is a subtle and ambient process; the audience rarely associates the output with its creator.

Tokuhisa et al. [9] presented the Rhythmism system that utilized two maraca-like devices for control of a live VJ performance. They argue that Rhythmism provided VJs with freedom of movement and allowed audiences to realize the power of the performance. Consequently, they conclude that manipulation should be an important part of the performance.

Nacenta et al. [5] and Cao and Balakrishnan [2] presented systems that track a laser pointer as an input device for remote interactions with a large display. While the laser pointer provides a very intuitive way to randomly access any portion of a wall-sized display, rotational jitter present in the hand can make it difficult to use for precise target acquisition. Moreover, ordinary laser pointers have only two degrees of freedom. This limits their use for complicated tasks such as resizing, rotating and layering of visual objects.

Figure. 2(a). Remote selection using breach gesture.

Figure. 2(b). Remote selection using squeeze gesture.

The VisionWand system [2] used simple computer vision to track the colored tips of a plastic wand to interact with large wall displays, from both close-up and a distance. A variety of postures and gestures are recognized by the system in order to perform an array of interactions in a picture application, such as selecting and scaling, query and undo.

Other systems use computer vision to track markerless hands using one or more cameras, with simple hand gestures for arms-reach interactions. For example, the Barehands system [8] uses hand tracking technology to transform any ordinary display into a touch-sensitive surface. Similarly, the Touchlight system [10] uses two cameras to detect hand gestures over a semi-transparent upright surface for applications such as face-to-face video conferencing or augmented reality. The major advantage of such vision-based techniques is their ability to track multiple fingers uniquely, which allows for more degrees of freedom when compared to standard input devices such as a mouse. However, this advantage of computer vision-based techniques has not yet been fully leveraged for interactions with large wall-sized displays. A major disadvantage of these systems is that they generally lack the real-time robustness required for VJ performances on location.

For this reason, researchers have designed systems that track specific markers on fingers and hands that achieve higher pointing precision. Gustafson et al. [4] discussed a scenario for tracking the user's hands to create screen-less spatial interactions. Zigelbaum et al. used the g-speak spatial operating environment [6] to create g-stalt, a gestural interface to control video media [11]. One of the disadvantages of g-speak is that it relies heavily on symbolic gestural input rather than simple multitouch gestures to perform remote pointing tasks. We addressed this issue in WaveForm, which was modeled on simple operations that are common in multitouch products, such as the iPad.

WaveForm Implementation

WaveForm uses a Vicon motion capture system to track markers on the location of the fingers and nose bridge of the VJ. The system uses this information to implement a perspective-based mapping of finger position to the remote screen [1]. In this perspective-based mapping, pointing a finger at the screen results in a touch event occurring on the apparent location at which the remote finger appears on the screen, from the point of view of the VJ.

To track the position of the nose bridge, the VJ wears a pair of glasses augmented with retro-reflective markers for tracking by the Vicon system. The location of the glasses is used to calculate the VJ's perspective relative to the plane of the display. The VJ also wears a pair of gloves with markers on the finger joints and tips. These are tracked by the Vicon system to determine the position of each individual finger. In the left hand glove, markers were placed on the index and ring finger, with a rigid arrangement of three markers at the back of the hand. The rigid arrangement was inversed in the right glove.

Our large display measures 1.65 m x 1.2 m, and is back-projected using a Toshiba X300 short-throw projector running at 1024 x 768 resolution. The display is augmented with retro-reflective markers to allow the software to determine its location relative to the VJ. The WaveForm software environment was written using WPF4.0, which is natively capable of tracking multi-

Figure. 3. VJ, with glasses and gloves, throwing a video from the iPad onto the large display using a Swipe gesture.

touch input events with minimal lag. This framework also supports high performance graphics for our use of video transparencies and blending.

The WaveForm software runs on a Windows 7 laptop that connects to the Vicon over Ethernet to obtain coordinates of the nose bridge and fingers. These are then mapped to a three dimensional model of the scene that includes the location of the display. The software then uses a perspective model of the display to determine the location of the fingers relative to the plane of the display.

The video palette was implemented on an iPad running a custom version of the Remote application, which communicates to the desktop running the WaveForm software environment over a TCP/IP connection.

WaveForm Gestures

We designed a suite of in-air gestures based on common multitouch interaction techniques that allow VJs to manipulate live video on a remote display. These in-air multi-touch gestures control the manipulation of videos on the large display:

1. Swipe

This gesture allows video content that is pre-cued on the iPad to be "thrown" onto the large display (see Figure 3).

2. Remote Touch

Our most basic technique consists of two *Remote Touch* gestures. They are used to implement virtual clicks in all the higher-level gestures in the gesture set. We developed two distinct gestures for generating remote touch events:

a) *Squeeze*. To perform the *Squeeze* gesture, the user points at the display using his index finger, and clenches his/her middle, ring and little finger. Clenching produces a touch-down event and unclenching generates a touch-up event in our WPF multitouch environment. This gesture was based on the idea of trying to reach and grab distant objects, and is similar to the "pistol" gesture used in g-speak [6].

b) *Breach*. The *breach* gesture was designed to facilitate multiple touch events. Here, the user points at the target using their index finger and moves their hand towards the screen. When the index finger crosses (breaches) a distance threshold away from the nose bridge, a touch-down event occurs in our WPF environment. When the index finger retreats to a position closer to the VJ than the distance threshold, it generates a touch-up event. This selection method more closely resembles touching a virtual screen at a fixed distance, within arm's reach, between the user and the display.

3. Translate

To perform a simple translation of a video object, the VJ the selects the video using a breach gesture, with their dominant hand. While maintaining this breach, the user then moves the dominant hand parallel to the display.

4. *Rotate & Scale*

This is a bimanual gesture in which the VJ uses *breach* gestures to select content. First, the video object is selected by moving both hands through the breach threshold. Moving one or both hands in a circular path then causes the video object to rotate on the screen. The video scales down when the user moves the hands

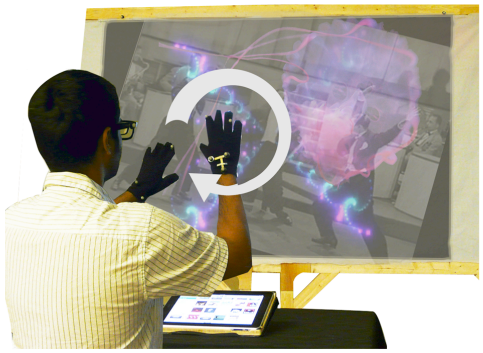

Figure. 4. VJ performing the Seek gesture with the large display.

closer together. Moving the hands apart causes the video to increase in size, while moving the hands simultaneously causes the object to translate.

5. *Layer Control*

This is a bimanual gesture that allows the VJ to control the layering of video objects on the display. When videos are brought onto the screen, they are overlaid on top of each other, as a stack. This ordering has implications for how the video content blends. It also affects selection, as only the top video object can be selected using simple *Remote Touch* gestures. To control in what layer a video object appears, the VJ can 'push' content to a lower order in the stack. To begin the process, the VJ selects the video with a *squeeze* gesture using their non-dominant hand, maintaining this selection. Upon selection, the relative distance of their dominant hand to the non-dominant hand, on the z axis (towards the screen), represents the absolute z location of the selected video within the layers of video objects on the display. When the user moves their dominant hand towards the display, away from their non-dominant hand, the video object is moved "away" from the VJ, below the videos in the layers below it. The breach threshold represents the bottom layer on the display. By bringing their dominant hand towards their face, away from their non-dominant hand, the video object moves toward the VJ, above other video objects in the stack. Releasing the non-dominant squeeze ends the gesture. This gesture is unique in that it allows for video objects to be selected in bottom layers. This is done by positioning the dominant hand towards the breach threshold prior to performing the *squeeze* gesture.

6. *Transparency Control*

To have finer control over the video mixing, the VJ can modify the level of transparency of a video using a bimanual gesture that is opposite to the *Layer Control* gesture. To begin the process, the VJ selects the video with a *squeeze* gesture using their dominant hand, maintaining this selection. Upon selection, the relative distance of their non-dominant hand to the dominant hand, on the z-axis (towards the screen), represents the *relative* transparency of the selected video. When the user moves their non-dominant hand towards the display, away from their dominant hand, the transparency of the video decreases. By bringing their non-dominant hand towards their face, away from their dominant hand, the transparency can be increased. Positioning the hands next to each other produces no change in the video object. Releasing the dominant hand *squeeze* ends the gesture.

7. *Seek*

In addition to manipulating the visual properties of the video, the user can also scratch through the video content of the objects on the display. This is one of the most important means of expression of a VJ, and can be connected to a scratching of the currently playing music file. To begin, the VJ selects the video with a *breach* gesture using their non-dominant hand. Moving the dominant hand in a circular motion, within the same plane of the display, scratches through the video. Clockwise rotation moves the playhead forwards, while counter-clockwise motion moves the playhead in reverse. This action represents the act of 'scratching' the video, in a similar fashion as a DJ would scratch a record. An additional advantage of using a circular gesture is that it does not have a distance limitation.

Discussion

Initial experiences with WaveForm suggest that audiences embrace having a VJ as a performance act on a podium, rather than standing behind a table with a laptop. While our system requires some training on behalf of the VJ, the gesture set appears well suited physically to the needs of the VJ. Further, we would like to investigate the responsiveness and accuracy of the design in a real world setting and ascertain audience and artist response. In future systems, we intend to support collaboration with a DJ and expand the gesture set with more advanced video manipulations, such as the ability to animate or morph the trajectory and shape of video images.

Conclusions

In this paper, we presented WaveForm, a system that allows VJs to directly manipulate video content on a large display on a stage, from a distance. WaveForm implements in-air multi-touch gestures to layer, blend, scale, rotate, and position video content on the large display. Initial observations suggest this leads to a more immersive experience for the VJ user, as well as for the audience witnessing the VJ's performance during a live concert. Rather than standing behind a laptop on a table, the system allows VJs to perform on stage, interacting with physical gestures that manipulate content on the audience display remotely. The VJ movements also serve as a performance vehicle that links the visuals on the display to the dynamical body image of the VJ in a way that allows the audience to appreciate VJing as an expressive bodily performance act rather than a multimedia experience.

References

[1] Banerjee, A., Burstyn, J., Girouard, A., and Vertegaal, R. Remote Multitouch : Comparing Laser and Touch as Remote Inputs for Large Display Interactions. *In Submission to GI 2011*, (2011).

[2] Cao, X. and Balakrishnan, R. VisionWand: interaction techniques for large displays using a passive wand tracked in 3D. In *Proc. UIST 2003*, 173.

[3] Engström, A., Esbjörnsson, M., and Juhlin, O. Mobile collaborative live video mixing. In *International Conference on Human-Computer Interaction with Mobile Devices and Services*, (2008), 157-166.

[4] Gustafson, S., Bierwirth, D., and Baudisch, P. Imaginary Interfaces: Spatial Interaction with Empty Hands and without Visual Feedback. In *Proc. UIST'10*, (2010), 3-12.

[5] Nacenta, M.A., Sakurai, S., Yamaguchi, T., et al. E-conic : a Perspective-Aware Interface for Multi-Display Environments. *Proc. UIST'07*, (2007), 279-288.

[6] Oblong Industries. http://www.oblong.com/.

[7] Retina Funk. http://www.flickr.com/photos/retinafunk/85772964/in/photostream/

[8] Ringel, M., Berg, H., Jin, Y., and Winograd, T. Barehands: implement-free interaction with a wall-mounted display. In *Ext. Abstract CHI 2001,* 367-368.

[9] Tokuhisa, S. dangkang, Iwata, Y., and Inakage, M. Rhythmism: a VJ performance system with maracas based devices. *ACM Int. Conference Proceeding Series; Vol. 203*, (2007), 204.

[10] Wilson, A. D. Touchlight: an imaging touch screen and display for gesture-based interaction. In *Proc. ICMI 2004*, 69–76.

[11] Zigelbaum, J., Browning, A., Leithinger, D., Bau, O., and Ishii, H. G-stalt: a chirocentric, spatiotemporal, and telekinetic gestural interface. In *Proc. TEI 2010,* 261–264.