g-stalt: a chirocentric, spatiotemporal, and telekinetic gestural interface

Jamie Zigelbaum, Alan Browning, Daniel Leithinger, Olivier Bau*, and Hiroshi Ishii

Tangible Media Group, MIT Media Lab
Building E15, 20 Ames St.
Cambridge, Mass. 02139 USA
{zig, abrownin, daniell, ishii}@media.mit.edu

*InSitu, INRIA Saclay & LRI
Building 490 Univ. Paris-Sud
91405 Orsay Cedex, France
bau@lri.fr

ABSTRACT
In this paper we present g-stalt, a gestural interface for interacting with video. g-stalt is built upon the g-speak spatial operating environment (SOE) from Oblong Industries. The version of g-stalt presented here is realized as a three-dimensional graphical space filled with over 60 cartoons. These cartoons can be viewed and rearranged along with their metadata using a specialized gesture set. g-stalt is designed to be *chirocentric*, *spatiotemporal*, and *telekinetic*.

Author Keywords
Gesture, gestural interface, chirocentric, spatiotemporal, telekinetic, video, 3D, pinch, g-speak.

ACM Classification Keywords
H5.2. User Interfaces: input devices and strategies; interaction styles.

INTRODUCTION
Human beings have manipulated the physical world for thousands of years through the gateway of a powerful interface—the human hand. Over the past half century we have spent more and more of our time manipulating a new, less-physical world—the digital world of computers. In this world able-bodied humans still employ their hands as the fundamental interface although our new hands are augmented by electromechanical devices that translate their actions into digital space. The standard computer mouse in particular is one such device. It channels the three-dimensional hand into a zero-dimensional pointer and confines it within a two-dimensional plane. This "pointer-in-plane" configuration is the fundamental basis of the ubiquitous graphical user interface (GUI) which has been a powerful and far reaching innovation that brought spatiality to bear on the previously mostly abstract and symbolic domain of computing. The GUI has been sufficient for interacting with most computers, but now as pixels become cheaper and more human-to-human interaction takes place in the digital world there is a need to widen the interface bandwidth between human and machine.

We seek to restore the human hand to it's full potential for interaction as a articulate, three-dimensional tool capable of complex gestural interaction. Towards that end, in this paper we present our work developing the g-stalt gestural interface. g-stalt is a tool for interacting with video media that is based on the g-speak [9] spatial operating environment (SOE).

Figure 1. The g-stalt interface.

GESTURAL INTERACTION
As for the hands, without which all action would be crippled and enfeebled, it is scarcely possible to describe the variety of their motions, since they are almost as expressive as words. – Quintilian [10]

Although science, technology, and our understanding of both have advanced significantly since the days when Engelbart's team developed the mouse at SRI, the interface layer between humans and machines has changed little in comparison. The possibilities for expanding the bandwidth of this interface through a more complete use of the hand and physical senses is compelling and challenging.

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

In this work we attempt to increase the articulation power of the human user through the implementation of a gestural vocabulary and graphical feedback system. With g-stalt we limit the expressive power of the human body by focusing only on specific configurations of the hands. We developed the term *chirocentric* (hand-centric) for this form of gestural interaction meaning that it is focused on both entire hands rather than the finger tips (e.g. gestural interfaces implemented in GUIs or multitouch interfaces) or the entire body. We prefer this term to using the existing term *free-handed* which is confusing since it seems to imply a completely unencumbered hand even though much work in this space is done utilizing gloves.

The word *chriocentric* is referential of John Bulwer's *Chirologia* [3] and the reverend Gilbert Austin's work *Chironmia* written in 1806 [1] which remains one of the most complete classifications of gesture. These works, and other more recent studies of gesture such as Efron's [5], McNeill's [8], and Kendon's [7] largely focus on those gestures of the body that accompany spoken language– gestures that serve as an auxiliary channel of communication. One challenge for research in gestural interaction will be to create usable, articulate gestures that can convey information to the computer (and, importantly, other users in the same space) both accompanying speech and independently of speech. The work in gesture studies can serve to provide insight into areas including how humans use gestures with each other, the typologies of gesture, how to interpret gestures, and what kinds of gestures to use in an interface. In the future researchers may have to create new systems of thought in order to include the computer as a top-level partner in gestural interaction.

G-SPEAK

g-speak is a software and hardware platform that combines gestural input, networking, and graphical output systems into a unified spatial operating environment. The version of g-speak running g-stalt uses a Vicon motion capture system to track passive IR retroreflective dots that are arranged in unique patterns on plastic tags placed on the back of the hand, the thumb, index, and middle fingers of simple nylon gloves. Each tag is tracked at over 100 Hz and with sub-millimeter precision in a room-sized 3D volume.

G-STALT

The g-stalt gestural interface allows users to navigate and manipulate a three-dimensional graphical environment filled with video media. They can play videos, seek through them, re-order them according to their metadata, structure them in various dimensions, and move around the space with 4 degrees of freedom (3 of translation, and 1 of rotation in the transverse plane). The videos are displayed on large projection screens and metadata for the videos is arranged on the projection surface of a table (Figure 1).

Interaction Themes

In creating g-stalt we wanted to see if we could create a complex gesture set that incorporated (using McNeill's typology [8]) *metaphoric* gestures to instantly manipulate features of a computational environment (similarly to the use of hot keys in a GUI), *iconic* gestures (the telekinetic gestures described below), *deictic* gestures (pointing), and what could be interpreted as Cadoz's *ergotic* gestures (pinching to move) [3]. We were concerned that too many gestures might be difficult for users to learn and remember (see Charade for a good accounting of concerns such as these [2]) but at the same time we are intrigued by the possibility of creating new, virtuosic interfaces that require time to learn but enable much greater power once learned. We developed the following themes to guide our work.

Theme 1: chirocentric

Although there are many ways to gesture we chose to limit the gestures available in g-stalt to specific configurations of the hands and fingers in space. This constraint helps to simplify the possibilities for action in g-stalt and allowed us to integrate well with g-speak's existing functionality.

Theme 2: spatiotemporal

We wanted to base g-stalt as much upon real-world phenomena as possible following the guidelines of Reality-Based Interaction [4]. By rooting the interaction design in conventional phenomena such as inertia, persistence in space, and solid geometry we designed the actions in g-stalt to mimic the real world.

Theme 3: telekinetic

We are intrigued by the science fiction idea of telekinetic powers—the ability to move matter with one's mind. We realized that with a gestural interface we could create a type of body-mediated telekinesis. For the functions that have direct and plausible gestural associations we used the most relevant gestural mappings that we could come up with, such as pinching to move space. For functions that had no real world analogs we tried to develop metaphorical bindings that made sense. We used the idea of telekinesis to structure the interactions where the user manipulates the spatial position of multiple videos directly.

Gesture Set

Figures 2–21 show the gestures implemented in g-stalt. Of these gestures pinch, two-handed pinch, stop all, lock, unlock, play all, the telekinetic gestures, change spacing, and add a tag were created by us, the rest of the gestures were developed by Oblong prior to this work.

Navigating Space

Figures 2 and 3 illustrate the *pinching* gestures. By touching the tip of index and thumb together on either hand the user grabs hold of space and is then able to translate the graphical environment isotonically with their hand. When the user pinches with both hands she can now translate and rotate the space.

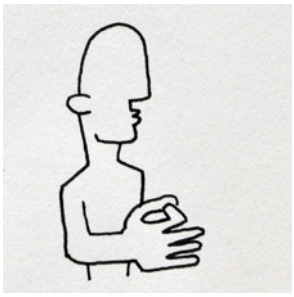

Figure 2. Pinch to translate though space.

Figure 3. 2-handed pinch for translation and rotation.

Figure 4. Stop all movement in space.

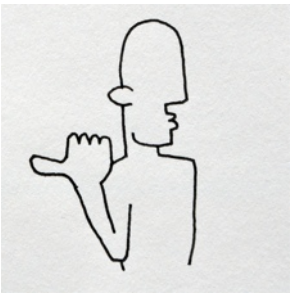

Figure 5. Reset space to the original view.

Figure 6. Point.

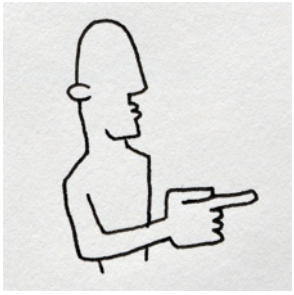

Figure 7. Click.

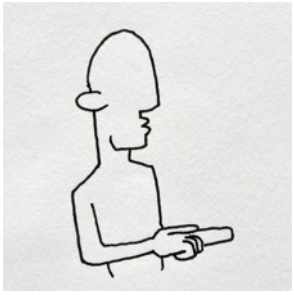

Figure 8. Lock. After clicking on a video.

Figure 9. Unlock. Must be made directly after lock.

Figure 10. Play. Can be combined with click.

Figure 11. Pause. Can be combined with click.

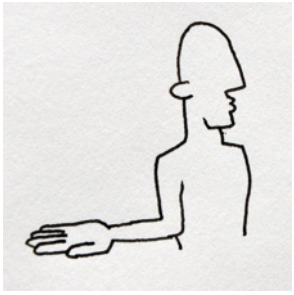

Figure 12. Reverse. Can be combined with click

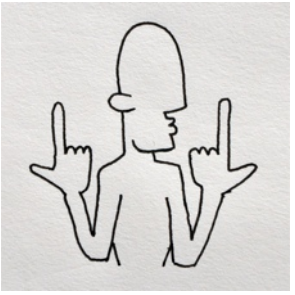

Figure 13. Play all the videos.

Figure 14. Telekinetic line creator in X axis.

Figure 15. Telekinetic line creator in Y axis.

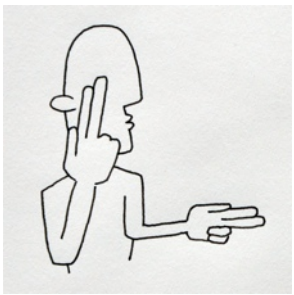

Figure 16. Telekinetic line creator in Z axis.

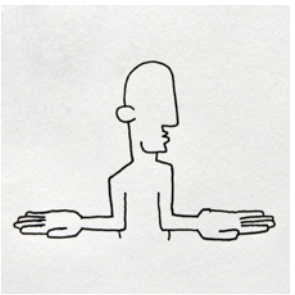

Figure 17. Stop all the videos.

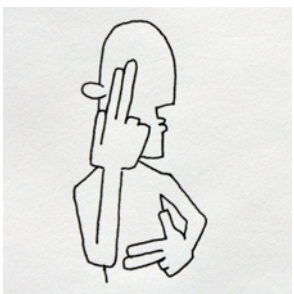

Figure 18. Telekinetic plane creator.

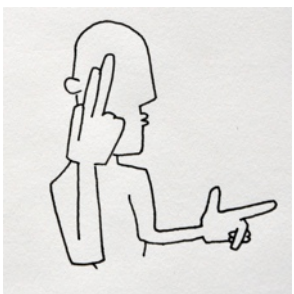

Figure 19. Telekinetic cube creator.

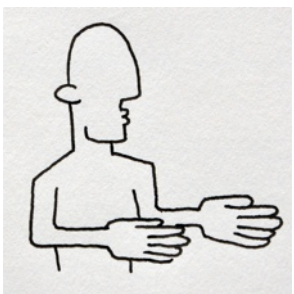

Figure 20. Change spacing between videos.

Figure 21. Add a tag to an axis.

Navigating Time

To play a video the user first *points* to the video (Figure 6) and then *clicks* (Figure 7) to zoom in on the desired video. Once zoomed-in the user can *play* (Figure 10), *pause* (Figure 11), and *reverse play* (Figure 12) the video with their other hand. They can also scrub through the video by *clicking* with their other hand and dragging it across the video's frame. When they release the initial clicking hand the video returns to its original position. If they make the *locking* gesture (Figure 8) after zooming in on a video the background fades out and the video is locked in the zoom position. When locked the user can manipulate the video with either hand. To unlock the video the user must make the *lock* gesture followed by the *unlock* gesture (Figure 9) which is the same as *click*—following the steps to lock a video in reverse.

Telekinetic Gestures

Beyond moving through space or time we wanted to allow the user to re-form the structure of objects in the space as easily and quickly as possible. To rearrange the spatial relationships of the videos in g-stalt the user touches one hand to their head and uses the other hand to define the shape of the structure they with to create (Figures 14, 15, 16, 18, and 19). The videos can be structured as a line along any of the three axes, as a 3D grid in the coronal plane, or as a cube. The direction that each video is facing does not change, only their position in space does.

Metadata

The videos used in g-stalt are classic American cartoons. We use every cartoon made by famous director Tex Avery during his employment at MGM studios—from *Blitz Wolf* in 1942 to *Cellbound* in 1955. While navigating space and time on the main screen the user can sort the videos by their metadata using the table surface in front of them. The tags Writer, Animator, Cast (voice actors), Character (featured cartoon characters), Duration (the duration of the cartoon in minutes), Month, Year, and Day (the date the cartoon was released) are available. By touching these tags on the table surface with an index finger the user picks up a tag. Then by touching that index finger to a finger on their opposite hand they can reorganize the space based on that tag. The target finger becomes a representation of the form of the space. If the videos are structured in a cube shape, the thumb represents the Y axis, the index finger represents the Z axis, and the middle finger represents the X axis (if you hold each of these digits orthogonally to each other they describe these three axes in space). To clear the tags the user can create a new telekinetic form, or touch their index finger to the finger holding the tag and then touch that index finger back against the table. It should be noted that the table is passive—g-speak identifies table touch events by the proximity of the fingers to the stored location of the table in space.

CRITIQUE

To date we have demonstrated g-stalt to over 250 people, many of these demonstrations took place during the MIT Media Lab's open house events. One of the main concerns that viewers had when first seeing g-stalt was that the gesture set was too complicated and would be difficult to learn. Chirocentric gestures are non-self revealing [2] making it difficult for new users to understand the possibilities for interaction.

We need to find the line between a gestural interface that is too simple (just pointing and clicking would not take real advantage of the hand's capabilities for expression) and one that is too complex. This balance will necessarily be impacted by the form of graphical feedback and interaction design used as well as the development of better techniques for learning and browsing gestures.

CONCLUSION

In this paper we have presented the g-stalt gestural interface. This work is part of our larger goal to create interfaces that privilege the expressive capabilities of the human body and that are rooted in existing human experience. Our goal is to remove the confines of the mouse from the hand, to re-enable the hand as a full-fledged citizen in our daily experience, and to shape our digital world around it. We remain far from achieving this grand vision. We hope that this work brings us a little bit closer.

REFERENCES

1. Austin, G. *Chironomia; or a treatise on rhetorical delivery*. Carbondale: Southern Illinois University Press. 1966. (Original work published 1806).
2. Baudel, T. and Beaudouin-Lafon, M. 1993. *Charade: remote control of objects using free-hand gestures*. Commun. ACM 36, 7 (Jul. 1993), 28-35.
3. Bulwer, J. *Chirologia, or, The natural language of the hand*. London: Thomas Harper. 1644.
4. Cadoz, C. *Le geste, canal de communication homme/machine. La communication instrumentale*. Technique et science informatiques. Volume 13, no. 1. 1994, pp. 31-61.
5. Efron, D. *Gesture and Environment*. King's Crown Press, N.Y., 1941.
6. Jacob, R. J., Girouard, A., Hirshfield, L. M., Horn, M. S., Shaer, O., Solovey, E. T., and Zigelbaum, J. 2008. *Reality-based interaction: a framework for post-WIMP interfaces*. CHI '08. ACM, New York, NY, 201-210.
7. Kendon, A. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press. 2004.
8. McNeill, D. *Hand and mind: what gestures reveal about thought*. University Of Chicago Press. 1992.
9. Oblong Industries. http://www.oblong.com/.
10. Quintilian, *Institutio Oratoria*. Loeb Classical Library Edition. XI, Chapter 3. 1990. (Original work c. 95 CE)