Evaluation of a Multimodal Interface for 3D Terrain Visualization

David M. Krum, Olugbenga Omoteso, William Ribarsky, Thad Starner, and Larry F. Hodges
{dkrum@cc, gte414w@prism, ribarsky@cc, starner@cc, hodges@cc}.gatech.edu

College of Computing, GVU Center, Georgia Institute of Technology, Atlanta, GA 30332-0280 USA

Abstract

Novel speech and/or gesture interfaces are candidates for use in future mobile or ubiquitous applications. This paper describes an evaluation of various interfaces for visual navigation of a whole Earth 3D terrain model. A mouse driven interface, a speech interface, a gesture interface, and a multimodal speech and gesture interface were used to navigate to targets placed at various points on the Earth. This study measured each participant's recall of target identity, order, and location as a measure of cognitive load. Timing information as well as a variety of subjective measures including discomfort and user preference were taken. While the familiar and mature mouse interface scored best by most measures, the speech interface also performed well. The gesture and multimodal interface suffered from weaknesses in the gesture modality. Weaknesses in the speech and multimodal modalities are identified and areas for improvement are discussed.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/Methodology, Input Devices and Strategies, Graphical User Interfaces, Voice I/O; I.4.9 [Image Processing and Computer Vision]: Applications—Gesture Recognition

Keywords: Multimodal interaction, evaluation, navigation, speech recognition, gesture recognition, virtual reality, mobile visualization, GIS.

1 Introduction

We believe that 3D visualizations will be among the mobile computing applications of the near future. Interaction with such mobile visualizations will be challenging. Users may be standing, moving, encumbered, or away from desktop surfaces. In any of these cases, traditional mouse and keyboard interfaces will be unavailable or unusable. Furthermore, the users may be attending to other tasks, spreading their cognitive resources thinly. We are investigating and characterizing candidate interfaces that could be used in mobile visualization. The interface must be unencumbering, expressive, and have low cognitive load.

In this paper, we evaluate a multimodal interface that might eventually be used in mobile visualization. This visualization might run on a wearable computer and be viewed through a headmounted display such as in Figure 1. The interface might also be used in an environment augmented with computers and large projected displays (Figure 2). In this instance, the user might interact with the visualization while standing at a distance from any display or keyboard, or attending to another task. Due to the nature of these displays, direct reference, such as with pen tap or finger touch, is difficult. For these reasons, we have focused on interaction modes and techniques that do not require direct reference to the display. Since we are doing an initial evaluation in a prototyping environment, we have eschewed making the system entirely mobile so that we can concentrate on the interface elements. However, the user can be away from the display and also enjoy some freedom of movement. Figure 3 is a diagram of the evaluation system.

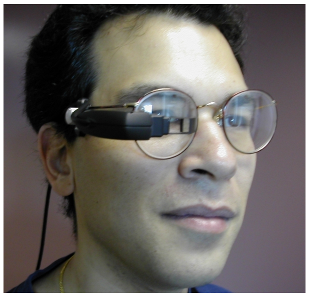


Figure 1: Wearable Computer Headmounted Display

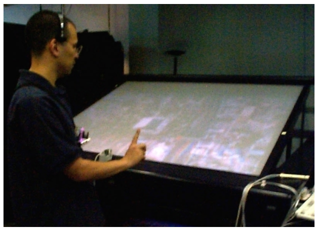


Figure 2: Large Projected Display

We are interested in mobile visualization and its interfaces because increasing compactness and increasing computing power is becoming available in wearable and other mobile systems. In addition, wireless networking and geo-located services (using GPS and other devices) allow mobile systems to access potentially unlimited data resources and inform that access with awareness of a user's location and context. Yet the ever smaller footprint of these devices and the fact that they will be carried and used everywhere makes the type of interface a critical issue. Without considering this issue, users may be faced with the prospect of having ever increasing resources at hand with less and less efficient ways of getting to them.

It is worthwhile to characterize and understand the inherent qualities of a speech and/or hand gesture interfaces rather than rejecting them in comparisons to more familiar and established interfaces. For many ubiquitous or mobile applications, these new interfaces may be the interfaces of choice because they best fit the environment and usage needs of the user. As mentioned earlier, in these applications, one may not have a mouse, keyboard, tracked 3D interaction device, or other wired device available. One may not have a desktop surface on which to operate. Furthermore, the user might stand apart from the display and computer or might be moving around. The user might also have her hands occupied either all or part of the time. Finally a speech and gesture interface with the appropriate affordances may not demand the attention or have the cognitive load of a traditional interface, which can be a key issue in many mobile visualization applications. For these and other reasons, it is worthwhile to study the hand gesture and speech multimodal interface in its own right to understand its characteristics. The issue is whether this interface performs effectively and accurately for its tasks, and if it does not, what characteristics need to be improved.

This paper builds on previous work[10], which provides a more detailed description of the interface architecture and implementation. The basic modes of interaction are hand gestures, captured by a camera worn on a user's chest, and speech recognition. These modes were used both separately and together for 3D navigation. This paper focuses on the formal evaluation of an initial multimodal interface in the context of a geospatial visualization system. This is the type of system that will be used in many location-aware applications. The extended navigation properties of the system provide a rich environment for testing the multimodal interface. In addition, a variety of other interface paradigms have been used with this system.

2 Related Work

There has been keen interest in multimodal control interfaces for a long period of time. Early work like Bolt's "Put That There"[3] has been followed by a large number of systems and studies. Some related work in multimodal interfaces and visualization is discussed below.

MSVT, the Multimodal Scientific Visualization Tool[7] is a semi-immersive visualization environment for exploring scientific data such as fluid flow simulations. The interface is composed of a pair of electro-magnetically tracked pinch gloves and voice recognition. Voice recognition provides over 20 commands and the gloves provide a variety of navigation, manipulation, and picking techniques. Visualization tools such as streamlines, rakes, and color planes are available. In our work we track hands without gloves, which encourages a more natural and unencumbered interaction. Furthermore, our visualization is a global terrain visualization with an extended range of scale, requiring richer navigation techniques.

Sharma et al.[14] describe another multimodal testbed composed of a virtual environment called MDScope and a graphical front-end called VMD. This system allows structural biologists to simulate the interaction of biomolecular structures. Interaction is through a simple command language composed of spoken actions executed with objects and parameters composed of both speech and gesture. The voice recognition system spots words from a continuous stream of speech while video streams from two fixed cameras are processed to yield 3D finger pointing and simple hand gestures. Our system uses a body mounted camera, so user mobility is enhanced.

BattleView[13] is a virtual reality battlefield application for supporting planning and decision making developed by the National Center for Supercomputing Applications. Much like the MDScope/VMD application, 3D pointing and simple hand gestures form the gesture part of the multimodal interface. IBM ViaVoice forms the speech recognition system. A multimodal integration module combines the recognizer streams. A state diagram describes the command language that allows users to navigate as well as select and manipulate virtual objects. Stereoscopic displays such as workbenches and single rear projected screens are supported. Again, a fixed single camera mounted on the display is used for gesture recognition, as opposed to a body mounted camera.

Several efforts have been directed towards multimodal pen and speech applications. The Multimodal Map[4] is a map based application, developed at SRI, that allows speech, handwriting, and pen gesture input. Various recognizers are managed in the Open Agent Architecture, a multi-agent framework. Quickset is a another 2D map application with a rich pen and speech interface developed at the Oregon Graduate Institute of Science and Technology[5]. Users can create and manipulate virtual entities on the map for a variety of applications, including medical informatics, military simulation and training, 3D terrain visualization, and disaster management. Quickset uses a 3 tier hierarchical recognition technique called Members-Teams-Committee. Member recognizers report results to one or more team leaders which apply various weighting schemes. These team leaders report to a committee which weights the results and provides a ranked list of multimodal interpretations.

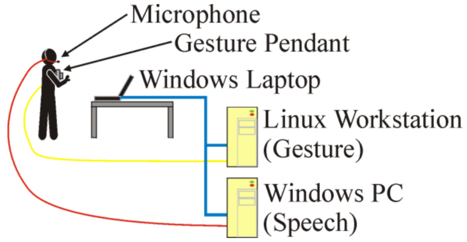


Figure 3: System Architecture

Quickset has also been adapted to Dragon[8], a battlefield visualization tool developed at the Naval Research Laboratory[6]. Features of the VR system include "digital ink" that is deposited on the 3D terrain surface by ray-casting. This ink plays the same role as pen strokes in 2D Quickset applications. Also, a 3D speech and 3D gesture vocabulary is integrated with the now available 3D information. An example would be the query "How high is this hill (3D gesture)?" Our multimodal interface is based on speech and hand gesture, rather than speech and pen stroke as in Quickset and Multimodal Map. Pen gestures require some reference or interaction with the display surface. With a body mounted camera, users can be distant from the display and still interact.

3 Method

This study explores four interfaces for navigation in a 3D visualization. These interfaces include a mouse interface, a speech interface, a hand gesture interface, and a multimodal speech and gesture interface. We have included the mouse interface as a baseline for comparison to help characterize the other interfaces. This study also attempts to determine the impact of each interface on cognitive load as well as take subjective measures such as discomfort and user preference.

3.1 Participants

Twenty-four students were recruited from an undergraduate computer game design course. The participants were male, and most had experience with 3D graphics in gaming or 3D design applications. Some had used commercially available speech recognition in the form of PC applications or telephone information systems. A small number had used applications with hand or arm gesture recognition. While not representative of the population in general, this group should be adaptable to new interfaces.

3.2 Apparatus

The apparatus used in this experiment consisted of a Pentium III 850MHz laptop running the VGIS visualization application. A Linux workstation ran vision algorithms for the gesture recognition

Movement Commands
Move {In, Out, Forwards, Backwards}
Move {Left, Right, Up, Down}
Move {Higher, Lower}
Speed Commands
Slower, Faster, Stop
Discrete Movement Commands
Jump {Forwards, Backwards}
Jump {Left, Right, Up, Down}
Jump {Higher, Lower}

Table 1: A Sample of Recognized Speech Commands

interface and sent packets with the results over a network to the laptop. A Windows NT system ran a speech recognition interface and also sent the results over a network to the laptop. A diagram of the system is in Figure 3.

3.2.1 VGIS

VGIS[11] is a 3D global geospatial visualization system that displays phototextures of the Earth's surface overlaid on 3D elevation data. Three dimensional models of buildings are also included for some urban areas. Recently, we have also included real-time 3D weather visualization in the VGIS framework. A hierarchical data organization allows the display of appropriate levels of detail and real time navigation of multiple gigabyte data sets.

VGIS supports a variety of 3D stereoscopic displays and interface devices such as mice, spaceballs, and Polhemus trackers. VGIS also supports a variety of navigation modes such as a downward-looking orbital mode, a helicopter-like fly mode, and a ground-following walk mode. A variety of configuration options and navigation commands are available.

While we have demonstrated each interface type (mouse, speech, gesture, and multimodal) in all of VGIS' navigation modes, we used a simplified navigation mode in this experiment. This was done to minimize the complexity, training, and time involved in this evaluation. Each subject was only scheduled for an hour block of time. The interface was limited to the downward-looking orbital mode. This navigation mode was further simplified by restricting roll, pitch, and yaw. Users could pan horizontally or vertically, and zoom in and out.

3.2.2 Mouse Interface

The simplified mouse interface uses a three-button mouse. Clicking the left button and dragging allows the user to pan horizontally and vertically. Pressing the middle button zooms in and pressing the right button zooms out. An additional zoom characteristic was that the mouse position determined the center of the zoom in and zoom out motions. This allows users to pan a small amount while zooming, allowing fine adjustments of their trajectories.

3.2.3 Speech Interface

The speech interface uses Microsoft's Speech API for recognition. No user training is needed, but some users with certain US regional dialects or non-US accents experience more recognition difficulties. Fortunately, synonyms are available for commands that often cause difficulty.

The speech interface provides three classes of commands (Table 1). There are movement commands that start the user moving in a particular direction. For example, the user can “Move left” or “Move right” to pan horizontally. “Move up” and “Move down” are used to pan vertically. A second movement command stops the previous movement and begins a new motion. This constraint was added after initial testing when we found that combined movements proved more difficult for users to control. The speed control commands, “Faster”, “Slower”, and “Stop”, allow the user to modify speed once a movement command has been given. The final class of commands, the discrete movement commands, “Jump left”, “Jump up”, “Jump down”, are much like the movement commands, except the user moves in small jumps without control of speed.

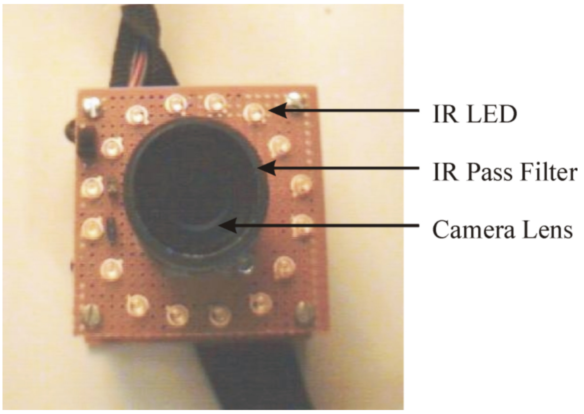


Figure 4: Gesture Pendant

3.2.4 Gesture Interface

The gesture interface uses the Gesture Pendant[1, 15]. It consists of a small, black and white, NTSC video camera that is worn on the user's chest (Figure 4). Since bare human skin is very reflective to infrared light, regardless of skin tone, an array of infrared emitting LED's is used to illuminate the camera's field of view. An infrared filter over the camera's lens prevents other light sources from interfering with segmentation of the user's hand. The limited range of the LED's prevents objects beyond a few feet from being seen by the camera. With a wide angle lens on the camera, the Gesture Pendant yields a field of view about 20 inches by 15 inches at a one foot distance. At that distance, although there is some fisheye distortion, a single pixel of the 320x240 video image should subtend around 1/16 inch.

The recognized gestures are shown in Figures 5 and 6. Sweeping a vertical finger in a horizontal direction allows horizontal panning. Sweeping a horizontal finger from the right hand up and down allows vertical panning. Sweeping a horizontal finger from the left hand up and down allows the user to zoom in and zoom out. A flat palm facing the chest stops any motion. As in the speech interface, a second movement command stops any previous movement and begins a new motion.

3.2.5 Multimodal Interface

The multimodal interface uses both speech commands and gestures. The speech component is basically the same as the speech interface; but with gestures used for rate control. For example, the user first gives a speech command such as “Move left”, which causes the motion in the left direction. The gesture component segments the user's finger tip and detects x and y motion of the finger tip. By moving the finger tip left and right, the user can speed, slow, or even slightly reverse the motion. Zooming and vertical panning are controlled by vertical displacement of a horizontal finger tip. Two additional speech commands were also added to provide alternative

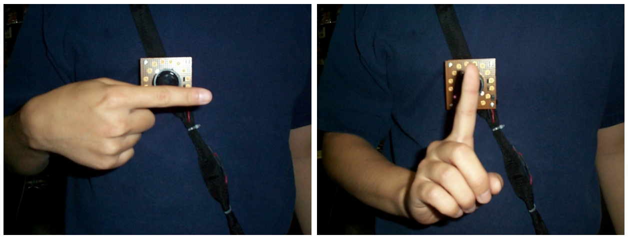


Figure 5: Moving the right index finger up and down causes vertical panning. Moving a vertical index finger left and right causes horizontal panning.

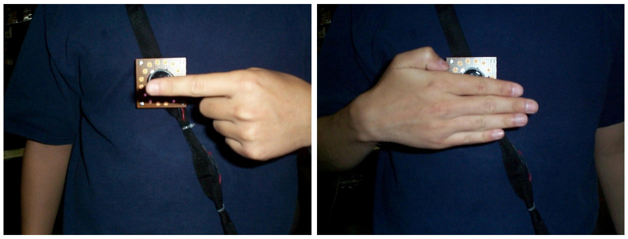
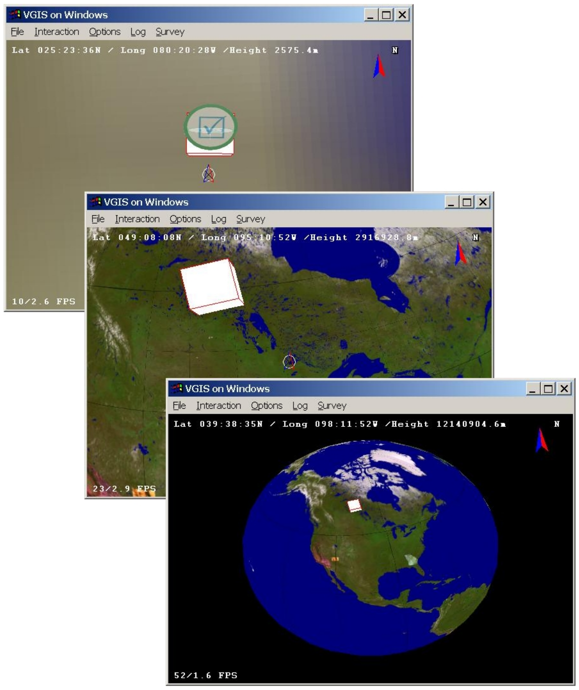


Figure 6: Moving the left index finger up and down causes zooming. An open palm stops movement.



VGIS on Windows
File Interaction Options Log Survey
N
VGIS on Windows
File Interaction Options Log Survey
N
VGIS on Windows
File Interaction Options Log Survey
N

```
Lat 025:23:36N / Long 080:20:28W /Height 2575.4m
10/2.6 FPS
Lat 049:08:08N / Long 095:10:52W /Height 2916928.8m
23/2.9 FPS
Lat 039:38:35N / Long 098:11:52W /Height 12140904.6m
52/1.6 FPS
```

Figure 7: Sequence of Images from an Experimental Task

commands for a few functions. “Horizontal” allows the horizontal finger tip displacement to determine both the direction and speed of horizontal panning and “Vertical” allows vertical finger tip displacement to do the same for vertical panning.

3.3 Design

The experiment compared the effect of a single variable (interface type) on a variety of objective and subjective measures. This experiment used a “within subjects” design, meaning that each participant used each and every interface type. The interfaces were presented to each participant in a unique order to counter learning effects.

A single interface task consisted of navigating to four different targets. These targets were each associated with a unique symbol. This task was repeated, with different target symbols and locations, for each of the four interfaces. There were two objective measures taken. The time needed to reach each target was measured. Participants were also given a memory test to determine if they remembered the symbols they saw, where the symbols were located, and in what order the symbols were encountered. This memory test was a tool to assess the cognitive load of the interface. One widely used result of cognitive psychology shows that there are severe limitations on working memory capacity[12]. Furthermore, when individuals are forced to use working memory or other cognitive resources, information is lost or displaced[2]. The cognitive load of a particular interface should be reflected in the quantity of information that an individual can remember while using that interface.

After each interface task, participants were asked to rate the interface for ten specific interface characteristics on five point disagree-agree response scales. They were also asked to write open-ended comments on aspects of the interface that were helpful and aspects that were problems.

At the end of the experiment, after experiencing each interface, participants were given the same ten interface characteristics and asked to order the interfaces by how well each interface expressed each characteristic. They were again given the opportunity to write open-ended comments on what was helpful or problematic for each interface and how the interface might be improved.

4 Procedure

Each of the twenty-four participants was given a consent form to read and sign. A questionnaire was given to each user to collect basic demographic information and assess their experience with computers, 3D graphics, speech recognition, and gesture interfaces. Participants were then shown a set of thirty symbols and asked to assign each a simple one word name. This allowed participants to become familiar with the set of symbols they would see during the task.

Participants were given several minutes to become familiar with each interface before starting the task. For interfaces involving speech recognition, they read the command list to ensure that they were familiar with all commands and the speech recognition process was working properly. They were allowed to try all commands and also practice navigation by finding and zooming in on Lake Okeechobee in Florida.

Participants were informed of the nature of the interface task and told to pay attention to symbols, location, and order of presentation. Participants began in a stationary position about twelve thousand kilometers above North America (see Figure 7). When an interface task began, a white cube appeared at a location in North America. As participants navigated closer and zoomed in, the white cube began to shrink. Eventually, the cube revealed a disc with a symbol. When the participant came to within about 4 kilometers, a chime sounded, signaling that the user had come close enough and should zoom out to find the next target. After four targets, a different chime sounded, signifying the end of the task. Participants were then given the memory recall test and after that, the post-task questionnaire. After all four tasks, the final post-experiment questionnaire was given.

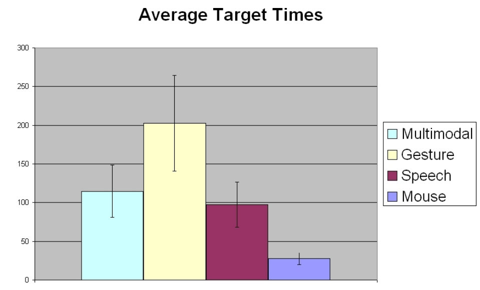


Figure 8: Average Target Time in Seconds for Each Interface

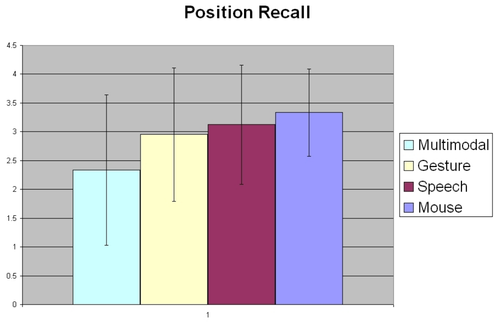


Figure 9: Average Number of Correctly Recalled Positions for Each Interface

5 Results

The experiment's objective measures were analyzed by computing a oneway ANOVA statistic [9], which examines variance in the dataset. This determines if any statistically significant conditions exist in a dataset. A Tukey post hoc analysis [9] was also performed, which examines variance between any two particular experiment conditions. This allows the particular statistically significant conditions to be identified.

The statistical analysis of the objective results shows significant differences in average target time ($p = 0.001$). The average target times of all of the interfaces were significantly different with the exception of the speech interface and multimodal interface. The mouse interface is significantly faster than the others. These results are illustrated in Figure 8.

The statistical analysis also shows a significant difference in recall of the target locations ($p = 0.013$). The mouse interface and multimodal interface were significantly different. However, the other interfaces had no significant differences. Furthermore, no significant differences among the interfaces were found at the ($p < 0.05$) level for symbol recall or order recall.

Participants were also questioned about ten interface characteristics on post-task and post-experiment questionnaires. The results of both questionnaires are very consistent, although the post-task questions were on a five point disagree-agree scale and the post-experiment questions asked respondents to rank the interfaces. The results of the five point scale questions are illustrated in this paper. The mapping of the responses were as follows (-2 Disagree, -1 Agree, 0 Indifferent, 1 Agree, 2 Strongly Agree). An ANOVA and Tukey post hoc analysis was performed to determine if the mean responses significantly differed between interfaces.

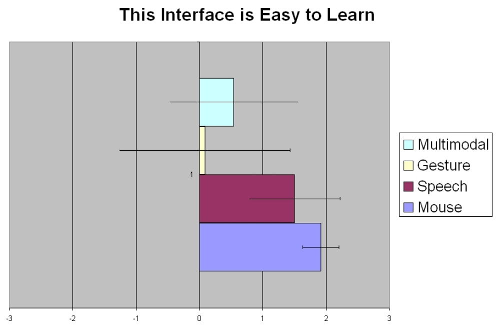


Figure 10: Response to Ease of Learning of Each Interface

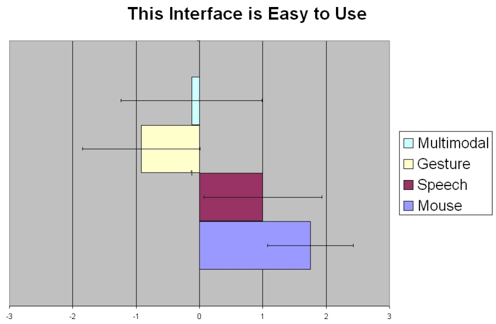


Figure 11: Response to Ease of Use of Each Interface

5.1 Ease of Learning

For the ease of learning characteristic (Figure 10), the interfaces fell into two groups. The participants felt that multimodal and gesture interfaces were not as easy to learn as speech and mouse. No significant differences were found between multimodal and gesture nor were there differences between speech and mouse. The users found the gesture component less easy to learn, either as a rate control for speech, or for motion control.

5.2 Ease of Use

Participants' responses for the ease of use question were significantly different for each interface. The ranking of the interfaces from easiest to hardest was mouse, speech, multimodal, and gesture (Figure 11). It is interesting to note that the speech interface had a positive rating while the multimodal interface had a neutral rating.

5.3 Errors

The speech and mouse interfaces were not significantly different in the participants' responses about error (Figure 12). However, the speech and mouse interfaces were better than the multimodal interface which was also better than the gesture interface. This suggests that users felt the speech interface to be similar to the mouse in error rate.

5.4 Speed

The participants' responses concerning the speed of the interfaces (Figure 13) perfectly reflected the objective measurements of average task time. The speech and multimodal interfaces were not

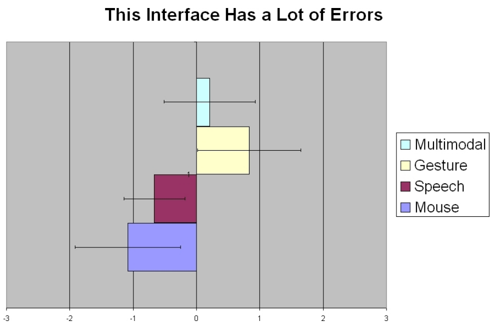


Figure 12: Response to Error Rate of Each Interface

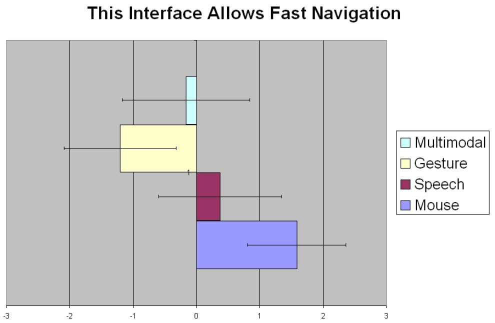


Figure 13: Response to Speed of Each Interface

statistically different. The mouse interface was felt to be fast and the gesture interface was felt to be slow.

5.5 Precision

The participants' evaluation of the precision of the interfaces paralleled their evaluation of the speed (Figure 14). Again, the speech and multimodal interfaces were not statistically different. The mouse interface was felt to be most precise and the gesture interface imprecise.

5.6 Cognitive Load

The multimodal interface was considered to provide the most interference to remembering the symbols (Figure 15). The mouse was evaluated as providing the least. This was also reflected in the location recall analysis. The gesture and speech interfaces did not significantly differ.

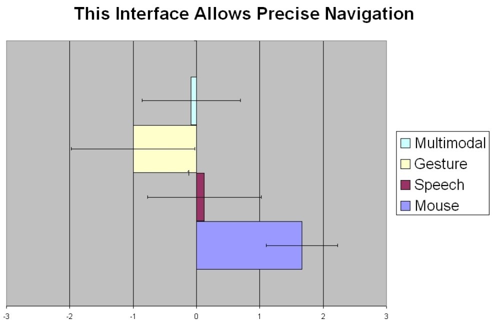


Figure 14: Response to Precision of Each Interface

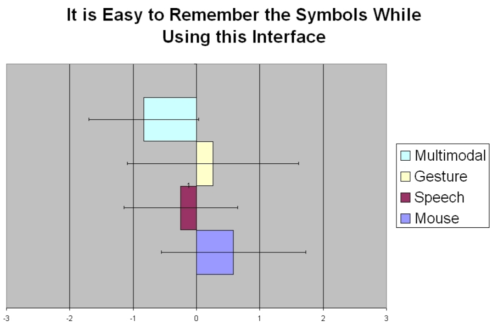


Figure 15: Response to Cognitive Load of Each Interface

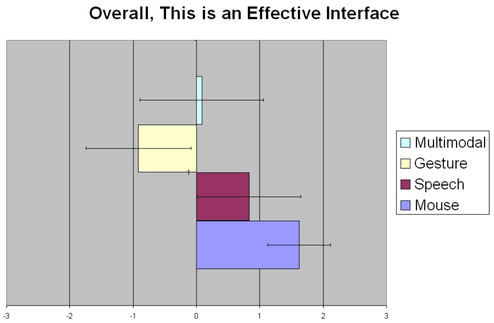


Figure 16: Response to Effectiveness of Each Interface

5.7 Effectiveness

Users strongly felt that the mouse interface was effective. Their responses for each of the interfaces were significantly different (Figure 16). The second highest support was for the speech interface followed by the multimodal interface and the gesture interface.

5.8 Presence

The participants were asked whether "This interface gives me the sensation of being in the map, i.e. I am present and part of the virtual environment." This was an attempt to determine if any of the interfaces improved the sense of presence in the visualization. However, there were no significant differences in opinion between the interfaces (Figure 17). This result is probably due to two factors. Presence was likely unaffected by interface choice. The environment did not seem to become more immersive with any of the interfaces. Secondly, presence is also a subtle concept to communicate. It is thus possible that the question was not clear to the respondents.

5.9 Comfort

The most comfortable interface appears to be the mouse interface followed by the speech interface. The multimodal and gesture interfaces appear to be the least comfortable to use. User responses did not distinguish the multimodal and gesture interfaces; they appear to be equally uncomfortable (Figure 18). Some respondents found it fatiguing to maintain their hand in front of the Gesture Pendant.

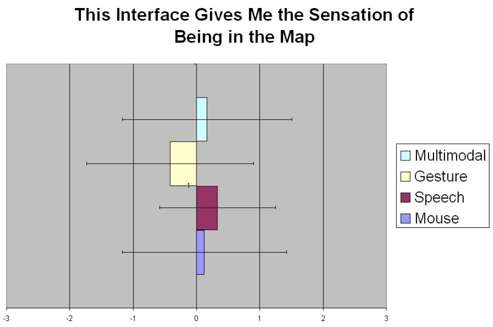


Figure 17: Response to Presence of Each Interface

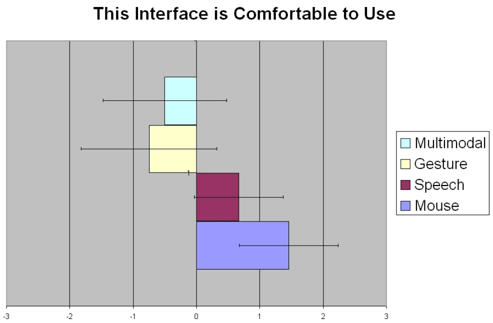


Figure 18: Response to Comfort of Each Interface

5.10 Desirability

After using and considering the characteristics of the interface, the participants were asked if they would like that interface on their own computers. The mouse was rated significantly higher than the other interfaces, but this reflects the status quo. The speech interface was second, but still significantly higher than the gesture and multimodal interfaces. The difference between attitudes towards the gesture and multimodal interfaces were not significantly different (Figure 19).

6 Conclusion

The familiarity of the mouse interface was one reason why the participants favored that interface. A few users were able to complete the navigation with the mouse so fast, they commented that it was difficult for them to recall targets. However, this concern was not widespread and was not reflected in the objective recall measures.

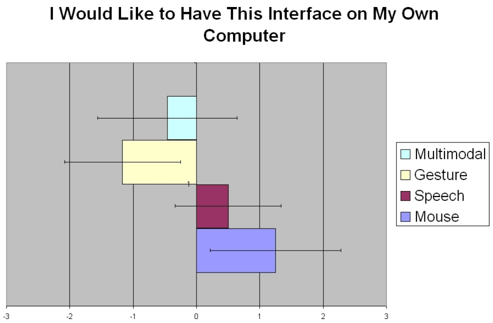


Figure 19: Response to Desirability of Each Interface

Overall, the speech interface was well regarded. The recognition lag in the speech interface was a source of difficulty for participants. Participants occasionally had to repeat commands and give some commands early to anticipate for lag. However, the participants' response to the speech interface was similar to the mouse for error rates and ease of learning. Precision was somewhat difficult, but users could adjust.

The gesture interface seemed to be the most difficult interface for the users. Errors in the recognition were a large source of problems. Precise movement was very difficult. Furthermore, some participants found it even uncomfortable to point a forefinger upward and move it left and right. Some wanted to use a thumb or point the forefinger down. Another difficulty with the gesture interface is that the wrist and fingers are held in a relatively static position with motion emanating from the arm. This unfortunately bypasses the fine motor control possible from the wrist and fingertips. It is this fine control that allows humans to pick up and manipulate small objects. This is showcased in such fast and precise activity as handwriting and typing. Perhaps gesture interfaces should focus on tapping into this set of motor skills and ability.

Since performing the task with the gesture interface took far more time than any of the other interfaces, and since participants were only expected to spend about an hour on the experiment, several participants did not complete the task for the gesture interface. However, this did not seem to greatly affect the results of this study.

The mouse and speech interfaces seem to rank highest by most measures. Of course, these interfaces are based on the most mature technologies. A few observations about the relatively low performance of the multimodal interface should be made.

While it is not surprising that the gesture interface was slowest and the mouse interface was the fastest, it is interesting to note that the speech and multimodal interfaces were not significantly different in speed. It was hoped that the additional expressiveness of the multimodal interface would have some benefit in speed. From the subjective results, it is apparent that the participants did not feel that the multimodal interface was more precise or faster than the speech interface. The addition of the gesture component did not improve performance. Furthermore, it hurt performance in some aspects. The multimodal interface was ranked most like the gesture interface in some subjective measures and indistinguishable from the gesture interface in ease of use, comfort, and desirability. The performance of the gesture component was certainly limited by the resolution of the video camera and the performance of the finger tip segmentation. A more robust and faster segmentation algorithm could significantly affect these results. While the Gesture Pendant was successful for home appliance control in [1], it appears that the navigation task in this study is of a different and more challenging nature. The navigation task requires far more gesturing for a longer period of time. It also requires a higher degree of precision and control over movement, so gesture timing is important.

For our objective of use in a ubiquitous or mobile visualization environment, where a mouse may not be available or handy, the results indicate that speech can be effective, at least for the extended navigation task presented. The results indicate that better gesture recognition is an important factor here and further work is needed to improve recognition. Furthermore, there may be different or more complicated tasks where the increased expressiveness of a multimodal interface would pay off. Different gestures should also be tried for improved comfort, ease of use, and precision.

7 Future Work

Future work would be to address the problems and limitations of the gesture interface. Both hardware and software enhancements are possible. Recognition might improve if the Gesture Pendant could capture and process 3D data. A highly detailed 3D image of the fingers could help take advantage of the fine motor control possible in the hand. This 3D imaging could be accomplished through a stereo camera pair. Depth information could be used to better segment the nearby hand silhouette from more distant infrared light sources and reflections off highly reflective objects. Depth information would also allow gestures along the Z axis and allow better differentiation of the wrist and finger tips. An alternative approach would be to use a single camera and a visible laser projected into a grid pattern. Measuring deformations in this structured light would allow 3D imaging of the hand. This would have the additional benefit of visibly illuminating the camera's field of view so users would know when their hand was visible to the camera. Also, this configuration could allow outdoor gesture use. While sunlight's broad spectrum and intensity can overwhelm the current Gesture Pendant's infrared illumination, the visible laser may be intense and narrow band enough for outdoor use. We are currently designing a structured light device.

Another line of work would investigate multimodal interfaces using speech and other pointer input devices. These devices would be less susceptible to recognition errors. However, mice are not very appropriate for our mobile and wearable applications. The head-mounted displays we are interested in are not amenable to stylus and touchscreen interaction. We are thus planning a user study to investigate speech and two handed input through IBM trackpoint devices mounted on rings or gloves. We feel that this interface might be very expressive, yet non-intrusive and non-encumbering enough for use in the mobile or wearable contexts we are investigating. Furthermore, the Twiddler, a chording keyboard often used with wearable computers, contains a trackpoint pointing device.

8 Acknowledgments

This work was supported by grants from the Department of Defense MURI program, administered by Army Research Office, and from the NSF Large Scientific and Software Data Visualization program. Invaluable technical assistance was provided by Daniel Ashbrook and Rob Melby.

References

[1] D. Ashbrook, J. Auxier, M. Gandy, and Thad Starner. Experiments in Interaction Between Wearable and Environmental Infrastructure Using the Gesture Pendant. In *Human Computer Interaction International Workshop on Wearable Computing (HCII2001)*, New Orleans, LA, August 2001.

[2] A. Baddeley. Working Memory. *Philosophical Transactions of the Royal Society London B*, 302:311–324, 1983.

[3] R.A. Bolt. Voice and Gesture at the Graphics Interface. *ACM Computer Graphics*, 14(3):262–270, 1980.

[4] A. Cheyer, L. Julia, and J. Martin. A Unified Framework for Constructing Multimodal Applications. *Conference on Cooperative Multimodal Communication (CMC98)*, pages 63–69, January 1998.

[5] P.R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal Interaction for Distributed Applications. *ACM International Multimedia Conference*, pages 31–40, 1997.

[6] P.R. Cohen, D. McGee, S. Oviatt, L. Wu, J. Clow, R. King, S. Julier, and L. Rosenblum. Multimodal Interaction for 2D and 3D Environments. *IEEE Computer Graphics and Applications*, 19(4):10–13, 1997.

[7] J.J. Laviola Jr. MSVT: A Virtual Reality-Based Multimodal Scientific Visualization Tool. *IASTED International Conference on Computer Graphics and Imaging*, pages 221–225, 1999.

[8] S. Julier, R. King, B. Colbert, J. Durbin, and L. Rosenblum. The Software Architecture of a Real-Time Battlefield Visualization Virtual Environment. *IEEE Virtual Reality*, pages 29–36, 1999.

[9] G. Keppel. *Design and Analysis: A Researcher's Handbook*. Prentice-Hall, Inc., 1991.

[10] D.M. Krum, O. Omoteso, W. Ribarsky, T. Starner, and L.F. Hodges. Speech and Gesture Control of a Whole Earth 3D Visualization Environment. *VisSym '02, Joint Eurographics - IEEE TCVG Symposium on Visualization*, May 27-29, 2002.

[11] P. Lindstrom, D. Koller, W. Ribarsky, L. Hodges, and N. Faust. An Integrated Global GIS and Visual Simulation System. *Report GIT-GVU-97-07*, 1997.

[12] G. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *The Psychological Review*, 63(2):81–97, 1956.

[13] V.I. Pavlović, G.A. Berry, and T.S. Huang. A Multimodal Human-Computer Interface for the Control of a Virtual Environment. *American Association for Artificial Intelligence 1998 Spring Symposium on Intelligent Environments*, 1998.

[14] R. Sharma, T.S. Huang, V.I. Pavlović, Y. Zhao, Z. Lo, S. Chu, K. Schulten, A. Dalke, J. Phillips, M. Zeller, and W. Humphrey. Speech/Gesture Interface to a Visual Computing Environment for Molecular Biologists. *International Conference on Pattern Recognition (ICPR)*, pages 964–968, 1996.

[15] T. Starner, J. Auxier, D. Ashbrook, and M. Gandy. The Gesture Pendant: A Self-illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring. *The Fourth International Symposium on Wearable Computers*, pages 87–94, 2000.