Speech and Gesture Multimodal Control of a Whole Earth 3D Visualization Environment

David M. Krum, Olugbenga Omoteso, William Ribarsky, Thad Starner, Larry F. Hodges

College of Computing, GVU Center, Georgia Institute of Technology, Atlanta, GA 30332-0280 USA
{dkrum@cc, gte414w@prism, ribarsky@cc, starner@cc, hodges@cc}.gatech.edu

Abstract

A growing body of research shows several advantages to multimodal interfaces including increased expressiveness, flexibility, and user freedom. This paper investigates the design of such an interface that integrates speech and hand gestures. The interface has the additional property of operating relative to the user and can be used while the user is in motion or standing at a distance from the computer display. The paper then describes an implementation of the multimodal interface for a whole Earth 3D visualization which presents navigation interface challenges due to the large magnitude of scale and extended spaces that are available. The characteristics of the multimodal interface are examined, such as speed, recognizability of gestures, ease and accuracy of use, and learnability under likely conditions of use. This implementation shows that such a multimodal interface can be effective in a real environment and sets some parameters for the design and use of such interfaces.

1. Introduction

Multimodal interaction provides multiple classes or modalities of interaction to a user. An early example is Bolt's "Put That There"[3] which integrated speech recognition and pointing gestures. Speech is a rich channel for human-to-human communication and promises to be a rich channel for human-to-computer communication. Gestures complement our speech in a number of ways, adding redundancy, emphasis, humor, and description. Multimodal interfaces crafted from speech and gesture have greater expressive power, flexibility, and convenience.

Multimodal interfaces can experience a decreased error rate, as compared to the unimodal component interfaces. This is partly due to the user's freedom to choose the means of expression. Since a large repertoire of expression is available, users will select and adapt to modes of expression that satisfy their preferences and minimize errors[13]. In noisy environments, the user can rely more on gesture or pen input. A user who is disabled or encumbered can use speech. Someone with a cold or an accent can employ more gesture or pen input. Multimodal interfaces also experience mutual disambiguation[14]. Recovery from some errors is possible because contextual information from the other modes allows the system to correctly re-interpret the user's intentions.

Multimodal systems appear to be a good match for spatio-visual applications, such as visualization and virtual reality. Gestures allow concise spatial references and descriptions. Speech allows rich command and query interactions. While tracked hand gestures have been used to navigate and interact in virtual environments for some time, these usually involve unwieldy tethered devices such as gloves. In general, gloves are cumbersome and imprecise in measuring hand orientation and posture[9]. They are also unwieldy to share with others. These, among other reasons, have led to work in vision based tracking devices.

For many wearable or mobile applications, one may not have a mouse, keyboard, tracked 3D interaction, or other similar input device. Furthermore, there may not be a desktop surface on which to operate. The user might stand a distance from the display or be moving around. The user may have her hands occupied either all or part of the time. It is worthwhile

Figure 1: *Orbital View from VGIS*

Figure 2: *Surface View from VGIS*

to understand the qualities and limitations of multimodal speech and gesture interfaces for particular tasks, rather than merely comparing performance with other interfaces.

In this paper we discuss parameters for a multimodal navigation interface and describe previous relevant work. We then discuss implemention of a multimodal navigation interface using speech and gesture for a whole Earth 3D visualization environment. This environment provides a rich set of interactions with several modes of navigation. We then evaluate interface characteristics such as ease of learning and use, gesture recognizability, system responsiveness, and navigation task performance.

1.1. The VGIS Environment

We have chosen the VGIS system[11] for the multimodal interface because it provides a broad set of 3D navigational tasks. VGIS is a whole Earth 3D terrain visualization that allows navigation through several magnitudes of scale. A user can travel from an orbital perspective of the entire globe, to a first person view of 3D building models and sub-meter resolution images of the Earth's surface (Figures 1 and 2). Navigation and paging of high resolution data occurs in real time and at interactive rates.

Navigating an extended 3D space such as VGIS is

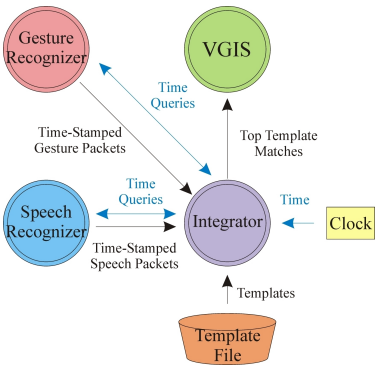

Figure 3: *System Processes*

complex due to the large magnitude of scales available. Wartell[20] cites three concerns for such applications:

1. Including scale, seven degrees of freedom must be managed.
2. In a virtual environment, good stereo imagery must be maintained.
3. Navigation methods must work at all spatial scales.

In the present work, we address concerns 1 and 3 with navigation constraints and aids that vary with scale. Interface design is further complicated by voice and gesture recognition engines that run on different machines and often have high error rates. We address these issues by collecting and integrating time stamped packets sent over a network by each recognizer.

2. Related Work

Our work differs from several gesture recognition projects such as Bimber's gesture recognition system[1, 2] which employed a tethered 6DOF tracker. We also employ a multimodal interface with speech recognition.

MSVT, the Multimodal Scientific Visualization Tool[10] is a semi-immersive scientific visualization environment that employs speech and gesture recognition, but uses electro-magnetically tracked pinch gloves. With the extended scale of our visualization, we require modified navigation techniques.

The MDScope/VMD system [18] for visualization of biomolecular structures and BattleView[16], a virtual reality battlefield visualization provide multimodal speech and gesture interaction. However, instead of fixed cameras as in these projects, our system uses a body mounted camera, so user mobility is enhanced.

Quickset is a 2D map application with a pen and speech interface[6] that has also been adapted to

Figure 4: *Gesture Pendant*

Dragon[8], a 3D battlefield visualization tool[7]. Our multimodal interface is based on speech and hand gesture, rather than speech and pen stroke in Quickset or speech and raycast strokes as in Dragon. The pen and stroke gestures require reference to a display surface. With a body mounted camera, users interact at a distance from the display.

Modes of Navigation
Orbit, Fly, Walk
Continuous Movement
Move {In, Out, Forwards, Backwards}
Move {Left, Right, Up, Down}
Move {Higher, Lower}
Discrete Movement
Jump {Forwards, Backwards}
Jump {Left, Right, Up, Down}
Jump {Higher, Lower}
Direction
Turn {Left,Right}
Pitch {Up, Down}
Speed
Slower, Faster, Stop

Table 1: *A Sample of Recognized Speech Commands*

3. Implementation

The multimodal interface was used on a variety of displays including a desktop Windows 2000 PC, an IBM laptop, and a Fakespace Virtual Workbench powered by an SGI Onyx2. Figure 5 shows some of these interfaces in use. Gestures were recognized by a Gesture Pendant[19]. Speech utterances were recognized by IBM ViaVoice. Speech and gestures were integrated with a late fusion method, as described in [14], where outputs of single mode recognizers are combined, as opposed to early fusion which uses a single recognizer to extract and integrate features from all interaction channels. Figure 3 is a diagram of the system.

3.1. Voice and Gesture Recognition

Voice recognition was performed by IBM ViaVoice. When speech utterances are recognized, an application time-stamps and transfers the commands over the network. Sample voice commands are listed in Table 1.

The Gesture Pendant is 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 hand gestures in the camera's field of view. At a one foot distance from the lens, the field of view is about 20 inches by 15 inches. 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.

The Gesture Pendant provides body-centered interaction that is unconstrained by the need for a surface and does not need to be tethered by wires. Gestures are with respect to the body and thus the proprioceptive quality of the interaction is enhanced since the user has an innate sense of the relation and movement of body parts with respect to one another. Mine et al.[12] have used this quality to develop 3D interaction tools in a tethered, tracked environment. In our work, the proprioceptive quality of the gestures permits the user to gesture without looking and to have an innate understanding of the amount and direction of hand movement. Since the gesture is done with the hand alone, without the need to grasp or manipulate an object, the user can attend to other tasks with the hands, eyes, or head.

The video image is segmented into blob regions, based on preset thresholds. If the blob conforms to previously trained height, width, and motion parameters, a particular gesture is recognized. The recognized gestures are listed in Table 2. The software can also extract the x and y coordinates of the centroid of the hand, allowing the hand to act as a pointer. Time-stamped packets describing the recognized gestures are sent over the network to the integration software.

3.2. Command Integration and Execution

Integration of gestures and speech utterances is performed by a semantic and chronological template matching process. Since the recognition processes query this process for a common synchronized time, gesture and speech packets can be ordered in time. The templates allow for a flexible specification of the command language. A variety of synonyms can be

Vertical Finger Moving Left: Pan Left Vertical Finger Moving Right: Pan Right
Left Finger Moving Up: Zoom Out Left Finger Moving Down: Zoom In
Right Finger Moving Up: Pan Up Right Finger MOving Down: Pan Down
Open Palm: Stop

Table 2: *Recognized Gestures*

specified for particular commands. Voice and gestures can work in a complementary fashion, with a particular command given by voice and described or given parameters by gesture. The voice and gesture commands can also work separately, but in parallel, for example allowing motion control by gesture while inserting new objects by voice.

Navigation commands are designed so that users can effectively navigate at all scales. The panning gain factors for the x and y directions are functions that vary with square of the altitude. As the user navigates closer to the Earth's surface, more precise panning control is available. However, since rotation is independent of scale, no special gain factor is needed. Scaling is integrated with changes in altitude. This follows Wartell's[20] scale factor adjustment to maintain and object's distance relative to the user.

Three particular navigation modes are available: Orbital Mode, Walk Mode, and Fly Mode. Orbital Mode presents a 3rd person point of view that always looks down from above. In Walk Mode, users are constrained to a ground following altitude. Fly Mode presents helicopter-like flight.

4. Application and Results

To evaluate the performance and effectiveness of the multimodal gesture interface, we had a group of users employ the gesture interface on a laptop display environment. We collected quantitative and observational information on user performance of specific tasks. We also interviewed them to ascertain their opinions and observations on the multimodal experience. We used a simple interface, with voice commands to initiate movement, and hand centroid tracking to control rate of movement.

4.1. Metrics

Several general criteria have been suggested for evaluating navigation tasks[4] and gesture interfaces[17]. We have concentrated on a subset of these criteria.

1. Gesture recognizability and responsiveness: how accurately and quickly the system recognizes gestures and responds.
2. Speed: efficient task completion.
3. Accuracy: proximity to the desired target.
4. Ease of Learning: the ability of a novice user to use the technique.
5. Ease of Use: the cognitive load of the technique from the user's point of view.
6. User Comfort: physical discomfort, simulator sickness.

4.2. Preliminaries and the Navigation Tasks

Six users became familiar with the multimodal interface and then performed a navigation task. None of these subjects had used the interface before. The users first trained the speech system by reading one story and then reciting the set of commands used in the interface. Recognition errors were corrected. This allowed the user to become familiar with the commands and the system to become familiar with the user's pronunciation. The process took 15 to 20 minutes.

Each user was shown how to position their hands so that gestures could be seen by the Gesture Pendant. The hand gesture recognizer required no user specific training. Users then experimented with the interface for 15 minutes. After this learning period, the users were verbally given a specific task. The users began in an orbital position (Figure 1), moved west, and zoomed into the Grand Canyon. The users then zoomed out, moving east to Georgia and into downtown Atlanta. From downtown Atlanta, they traveled in fly mode to the Georgia Tech campus, switched to walk mode, and parked in front of Tech Tower (the main administration building). These navigation activities required several fine adjustments as the user neared each goal. Users employed most of the multimodal commands (if not all) in this task.

4.3. Recognizability and Responsiveness

Voice recognition lag was a factor in the performance of the users. Also, users would sometimes have to repeat commands. The hand centroid tracking performed better. This was aided by more immediate and direct visual feedback for the hand motion (e.g., a turning movement would immediately speed up, or slow down based on hand movement) in a continuous process.

Studies on several types of interfaces, including those used in virtual environments[5, 21], indicate that tasks require system responsiveness to be 0.1 seconds or less. The hand tracking fell in this range. However,

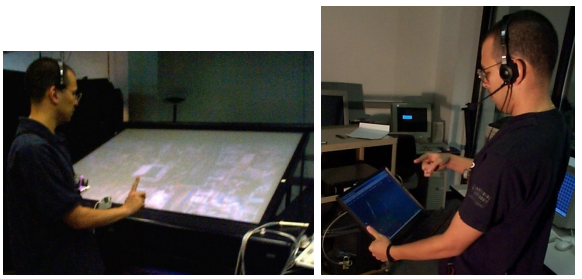

Figure 5: *Workbench and Mobile Interfaces*

voice recognition was slower. This mostly affected actions that required precise movements, such as when a user would try to position herself directly over a particular building. The multimodal interface with hand tracking was helpful in such actions. In future versions of this interface, we will be concentrating on two areas of improvement. We have already constrained the spoken word vocabulary and grammar of the recognizer, making recognition faster and more accurate. We will also be increasing the accuracy and precision of the gesture interface.

4.4. Performance on Navigation Tasks

The average time for task completion was 10.1 minutes with a standard deviation of 4.0 minutes. Each user gave between 50 and 100 spoken commands. The task with a mouse interface took about 3.5 minutes. The time for task completion in the multimodal interface was certainly affected by errors and delays in voice command recognition.

The accuracy of the navigation task was reasonably good with most errors occurring during adjustment of the more detailed movements. Again, this was mostly affected by delays or errors in voice recognition. Some users took the strategy of speaking a command ahead of time to allow for the delay. The hand gestures helped since one could slow or even stop a movement in preparation for a new voice command.

4.5. Ease of Learning, Ease of Use, Comfort

Users could remember both the voice and the gesture commands and some felt they were much easier to learn than keystroke commands. An important quality of the voice commands was that nearly every command had a mapping in all three modes. If particular commands work only in a certain mode, a user who tries a command in the "wrong" mode and fails may conclude that the command does not exist. An example is the "move down" command which changes altitude in Fly Mode, but tilts the user's view downward in Walk Mode. Further, several commands can map to the same action such as "move in" and "move forward."

Although some commands used different gesture mappings (upward finger movement increases rate of motion for "move higher" but decreases the rate of motion for "move lower"), there was not much confusion. The proprioceptive nature of the hand gestures made their interactions easier to remember. Furthermore, fast visual feedback informed users if they started moving in the opposite direction.

Some users desired gestures that did not require repositioning the hand for left/right and up/down gestures. This has been addressed with code to segment and track only the finger tip. Also, users would sometimes move their hand out of the camera's field of view. A cursor indicating hand position may address this problem. None of the users noted discomfort due to cybersickness. In some cases, there was some fatigue from holding the hand in front of the pendant.

5. Conclusions

While the Gesture Pendant is effective in many indoor environments, it is less effective outdoors. The sun's broad spectrum and intensity overwhelms the Gesture Pendant's infrared illumination. We are developing a new Gesture Pendant that uses a visible laser for structured light. The camera's field of view will be visibly illuminated and 3D imaging of the hand will be possible. The set of possible gestures should be significantly larger.

The multimodal interface proved easy to learn and effective in a navigation task that required many movements, including fine control, changes of mode, and navigation over an extended 3D space. The users had to plan and execute several commands to reach a target which was initially out of sight. Even under the increased cognitive load of this activity, users could successfully complete their task.

In the future, we will be conducting a series of a formal evaluations. The first user study has already begun and examines the cognitive load of various interfaces: multimodal, speech-only, gesture-only, and mouse. Preliminary results show clear benefits of the multimodal interface over the gesture-only interface.

6. Acknowledgments

This work was supported by grants from the DoD MURI program administered by ARO and from the NSF Large Scientific and Software Data Visualization program. Daniel Ashbrook and Rob Melby provided invaluable technical assistance.

References

1. O. Bimber. "Continuous 3D Gesture Recognition: A Fuzzy Logic Approach" *Fraunhofer Institute for Computer Graphics*. Report 98i013-FEGD (1998).

2. O. Bimber. "Gesture Controlled Object Interaction: A Virtual Table Case-Study" *Computer Graphics, Visualization, and Interactive Digital Media*, Vol. 1, Plzen, Czech Republic (1999).

3. R.A. Bolt. "Voice and Gesture at the Graphics Interface." *ACM Computer Graphics*, 14,3 pp. 262-270 (1980).

4. D. Bowman. "Interactive Techniques for Common Tasks in Immersive Virtual Environments: Design, Evaluation, and Application." PhD Thesis, Georgia Institute of Technology (1999).

5. S. Bryson "Implementing Virtual Reality." *SIGGRAPH 1993 Course #43 Notes*, 1.1.1-1.1.5; 16.1-16.12 (1993).

6. P.R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, J. Clow. "Quickset: Multimodal Interaction for Distributed Applications." *ACM International Multimedia Conference*, New York: ACM, pp. 31-40 (1997).

7. P.R. Cohen, D. McGee, S. Oviatt, L. Wu, J. Clow, R. King, S. Julier, L. Rosenblum. "Multimodal interaction for 2D and 3D environments." *IEEE Computer Graphics and Applications*, 19(4), pp. 10-13 (1997).

8. S. Julier, R. King, B. Colbert, J. Durbin, L. Rosenblum, "The Software Architecture of a Real-Time Battlefield Visualization Virtual Environment", *IEEE Virtual Reality*, Houston, Texas: IEEE Computer Society, pp. 29-36 (1999).

9. D. Kessler, L. Hodges, N. Walker. "Evaluation of the CyberGlove as a Whole-Hand Input Device." *ACM TOCHI*, 2(4), pp. 263-283 (1995).

10. J.J. Laviola, Jr. "MSVT: A Virtual Reality-Based Multimodal Scientific Visualization Tool." *IASTED International Conference on Computer Graphics and Imaging*, pp. 221-225 (1999).

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

12. M.R. Mine, F.P. Brooks, F.P., C.H. Sequin. "Moving Objects in Space: Exploiting Proprioception in Virtual-Environment", *SIGGRAPH 97*, pp. 19-26 (1997).

13. S.L. Oviatt, R. vanGent. "Error Resolution During Multimodal Human-Computer Interaction." *International Conference on Spoken Language Processing*, Vol. 2, 1996, University of Delaware, pp. 204-207 (1996).

14. S.L. Oviatt. "Mutual Disambiguation of Recognition Errors in a Multimodal Architecture." *ACM Conference on Human Factors in Computing Systems (CHI'99)*, Pittsburgh, PA, May 15-20, pp. 576-583 (1999).

15. S.L. Oviatt, P.R. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, D. Ferro. "Designing the User Interface for Multimodal Speech and Gesture Applications: State-of-the-Art Systems and Research Directions." *Human Computer Interaction*, Vol. 15, No. 4, pp. 263-322 (2000).

16. V.I. Pavlović, G.A. Berry, 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).

17. Y. Sato, M. Saito, H. Koike. "Real-Time Input of 3D Pose and Gestures of a User's Hand and Its Application for HCI." *IEEE Virtual Reality*, pp. 79-86 (2001).

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

19. T. Starner, J. Auxier, D. Ashbrook, M. Gandy. "The Gesture Pendant: A Self-illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring." *International Symposium on Wearable Computers*, Atlanta, GA: IEEE Computer Society, pp. 87-94 (2000).

20. Z. Wartell, W. Ribarsky, L. Hodges. "Third-Person Navigation of Whole-Planet Terrain in a Head-Tracked Stereoscopic Environment." *IEEE Virtual Reality*, Houston, TX: IEEE Computer Society, pp. 141-148 (1999).

21. B. Watson, N. Walker, W. Ribarsky, V. Spaulding. "The Effects of Variation of System Responsiveness on User Performance in Virtual Environments." *Human Factors*, Vol. 40, No. 3, pp. 403-414 (1998).