The Magic Wand

Jan Ciger[*]
Mario Gutierrez[†]
Frederic Vexo[‡]
Daniel Thalmann[§]

VRlab – EPFL
CH-1015 Lausanne
Switzerland

Abstract

We want to present a multimodal user interface for interaction with the virtual environment back–projected on the large projection screen. We use the interaction metaphor of a "spell–casting" wizard (the user) using a "magic wand" to interact with the VR environment and to complete some tasks. Our contribution is an user interface, which tries to take advantage of the past experience of the user such as fairy–tales or fantasy movies[1].

Keywords: multimodal, interface, non–obstructive, virtual reality, posture and speech recognition

1 Introduction

There are many approaches to the interaction within virtual environments. Most of them use additional hardware devices such as 3D mice, joysticks and different kinds of trackers. Various interaction metaphors are used, e.g. joysticks for locomotion, 3D mice for selecting and moving objects etc.

Problem with all these approaches, when used in the front of the large screen, is the learning curve, since they don't mirror any familiar "interface" from the real world for the regular users and/or obstructiveness e.g. full body motion capture requires many sensors attached to the user.

The other option is to use data–gloves or more complex interfaces such as haptic, which provide a more natural way to interact with the application. But these seem to be obstructive and problematic in practice. The usability tends to be limited due to lack of accuracy in data–gloves and very limited working volume of the haptic systems. They are not very well suited for the use with the projection screen. On top of it, these devices tend to be very expensive and hard to obtain.

We propose the idea of the "magic wand" as a very simple and natural interface to the virtual world. The user is in the role of the "magician" able to "cast spells" by pointing the wand and using voice commands recognized by the computer system.

The contribution of this paper is a non–obstructive (in the sense not restricting the user compared to e.g. magnetic motion tracking system), intuitive interface which tries to draw upon a priori knowledge of the user carried over from their childhood, e.g. fairy tales or fantasy movies, thus flattening the learning curve for interaction in the virtual environment.

2 Background

The idea of using the "magic wand" as an interaction metaphor is not new. Such device has been used more as a 3D mouse either to navigate through the environment or to grab and manipulate objects [8]. Implementations of such devices tend to mix the concept of the "magic wand" with less "traditional" control mechanisms such as buttons and joysticks e.g. [5] or [4], turning a simple interface into an unintuitive hi–tech artifact.

Natural language interfaces are more recent effort, enabled by progress in the speech recognition and understanding field. The idea of using speech input is not new, but only recently with the arrival of more robust speech recognition technologies and availability of the cheap computing power enabled the use of the continuous speech and speaker independent systems with deeper level of understanding of the spoken language. A lot of work was done especially for the military applications, using the speech as an advanced control mechanism [16] or [17], where the speech was used as a part of the multimodal interface.

The multimodal interfaces are an approach trying to merge several input (and output) modalities, such as speech, gestures, pen input or various devices. They enable the user to interact with the virtual environment in a similar way how he communicates in everyday life — for example "Move **that** box to the door!", where the box is

[*] jan.ciger@epfl.ch

[†] mario.gutierrez@epfl.ch

[‡] frederic.vexo@epfl.ch

[§] daniel.thalmann@epfl.ch

[1] The magic wand is a rather common artifact there

selected by hand gesture or pointing.

Probably the first multimodal application was the famous MIT's "Media room", described in the work of R. A. Bolt [2] from 1980, which implemented the "put that there" interface by tracking the directions of the user's hands and using a hardware based speech recognition system.

The work of Nijholt and Hulstijn [10] describes a multimodal interface to a virtual character (speech and keyboard input). Krum and Omoteso [7] make a comparison between the multimodal (gestures performed by the "gesture pendant" combined with speech) and classical (keystrokes) interfaces used in a GIS environment. They conclude, that actually many users found the multimodal interface much easier to use than the keystrokes.

In the paper titled "Ten Myths of Multimodal Interaction" [11], Oviatt summarizes the common problems encountered when designing the multimodal interface, e.g. assuming, that one modality is dominant or that the speech is the primary modality of the system, which includes it. Neither is true, one user may use the system as "Move the blue cube to the left" but the other may just drag the cube to the left directly, achieving the same goal. In the presented work, we tried to avoid these problems by using two interface modes complementing each other.

3 Interface Description

In virtual environments there are few basic interactions to be performed:

- navigation inside of the virtual world
- manipulation of objects
- communication with real or virtual characters

The "magic wand" provides two distinct modes which can be used to perform the first two tasks. One example implementation is described in the section 5.

We use two interface modes for the "magic wand" both accompanied by the speech recognition as a way to "cast spells" – i.e. invoke some actions (similar interface is described in [9]) :

- pointing mode
- posture recognition mode

One of the two modes is selected by the application according to the context.

The pointing mode is used to pick specific regions of the scene. The target is selected by pointing the wand at it and issuing the voice command (keyword) to confirm the selection. In this mode, the "magic wand" is used as a conventional pointer. We are not interested in the exact position of the wand, but instead in the point it is aiming at, in contrast to e.g. [5], where it is used as a 3D mouse (i.e. tracking the position mainly and using buttons to pick objects).

The posture recognition mode is one of the innovative aspects of our "magic wand", because it adds an additional processing layer, enabling more complex actions to be performed. In this case we are interpreting the orientation of the wand and trying to recognize the posture assumed by the user. Four basic postures are recognized by our interface:

- wand pointing left
- wand pointing right
- wand pointing forwards
- wand pointing backwards

Posture recognition reduces the complexity of the tracking data by classifying them into the small set of postures. By classifying the orientations, we are able to implement an intuitive mechanism for choice. Instead of adding a graphical interface, [2] we use the "magic wand" as a way to indicate the direction we wish to follow or to choose one of several alternatives.

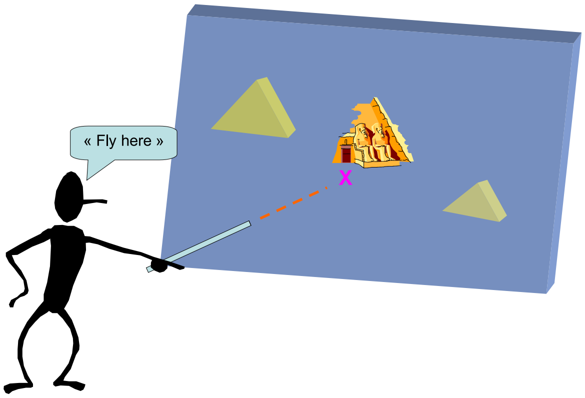

Figure 1: User interface

The postures are used for navigation in the virtual environment. But more complex interaction can be achieved as well. The "magic wand" can be used in combination with "spells" (keywords) to specify an additional information[3] for the object selected by voice ("spell").

We use the speech recognition engine just in a very simple manner, to recognize a small vocabulary of keywords, thus limiting the complexity of the problem and improving the robustness. There can be a lot of people using the "wand" and we want to achieve at least some level

[2] for instance, a set of buttons to select the directions or pick the available objects. This can lead to a less intuitive interaction and be out of context, reducing the immersion

[3] e.g. direction to move

of speaker independence. Otherwise we would have to require special enrollment procedure for each user, therefore making the learning process for a new user more difficult and the approach more tricky to use.

Figure 1 shows the user “casting a spell”, in our case asking the application to fly to the specified place in the virtual world.

4 System Architecture

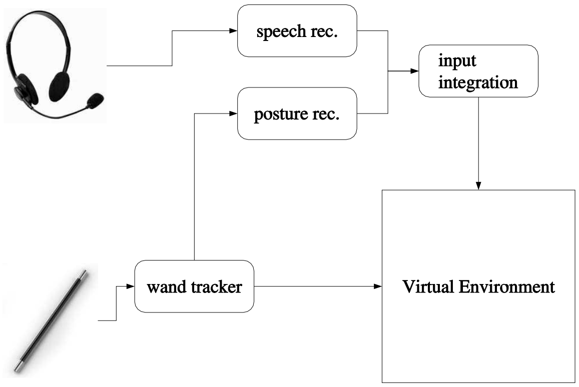

Figure 2: System Architecture

Figure 2 displays the overall architecture of our multimodal interface. The wand is being tracked by one magnetic sensor of the Ascension Flock of Birds (FOB) attached to it. The data from the wand (sensor) are copied into two streams, one stream is used for posture recognition and the second one is fed directly into the VR application to be used for camera management and pointing.

The data from the speech and posture recognition are combined at one place to enable the multimodal interaction e.g. selection of the objects by using the wand and manipulation of it by speech keywords. The aggregated, higher level actions are fed into the VR application for further processing and execution.

4.1 Tracking and Posture Recognition

This section describes the principles and methods used in the implementation of the tracking and posture recognition. The goals of this interface component are:

- To allow the user to indicate four basic postures (listed in section 3).
- To provide a way to point out any region or object in the VR environment.

We had to find a way to acquire (track) the orientation of the “magic wand” and process the data to identify the adopted posture. At the same time, we should use this information for the camera management and pointing mode.

Many different techniques have been used to solve the problem of posture recognition and orientation tracking of pointing devices (including the human hand, 3D mice, 3D wands and similar equipment) they can be roughly classified in two main categories:

- Optical tracking
- Sensor based tracking

Optical tracking techniques offer great benefit of a less obstructive interaction. In most cases, the user doesn’t need any special equipment, except maybe for some markers to simplify the tracking of the hand or the whole arm. Authors like Segen, et. al. [15] eliminate the use of any additional markers and try to directly track the user hands.

The main drawback of optical tracking is the lack of robustness. In general terms, special lightning conditions and/or special cameras are required to obtain useful results (e.g. [14]) and there is still problem with occlusions by the user’s body or parts of the scene.

Sensor based tracking has been widely used in VR applications, sometimes in combination with optical tracking as in [1]. Magnetic sensors are the most common alternative. They offer enough precision in the data acquisition and are rather robust to external conditions, providing that no metallic parts are in contact with the sensors and that no other sources of magnetic interference are present. The main drawback of this technology is the need to use wired sensors attached to the pointing device or to the user body. However, the strongest point of magnetic tracking is the ease of implementation. A magnetic tracking system such as the Ascension Flock of Birds delivers the orientation data in a practical and robust way.

We decided to implement the first prototype of the “magic wand” interface using the magnetic sensor–based tracking technology. It solves the data acquisition problem in a fast and reliable way.

The “magic wand” was thus implemented as a wooden wand with a magnetic sensor attached to it. The sensor delivers the information in the form of three rotation angles (Euler angles) which indicate the orientation of the sensor relative to the origin (the emitter).

The algorithm used to recognize the postures uses a very straightforward approach. We are essentially dividing the 3D space of angle rotations (measured by the magnetic sensor) into five regions (postures to recognize, plus neutral position), which are directly mapped to the basic postures we need to identify. Figure 3 shows how the 3D space has been segmented to map each zone to a specific posture.

We are able to sample the sensor’s orientation at 60Hz, but we use the average value of every two measures, getting a final sample rate of 30Hz. The posture recognition component provides a posture information about the wand

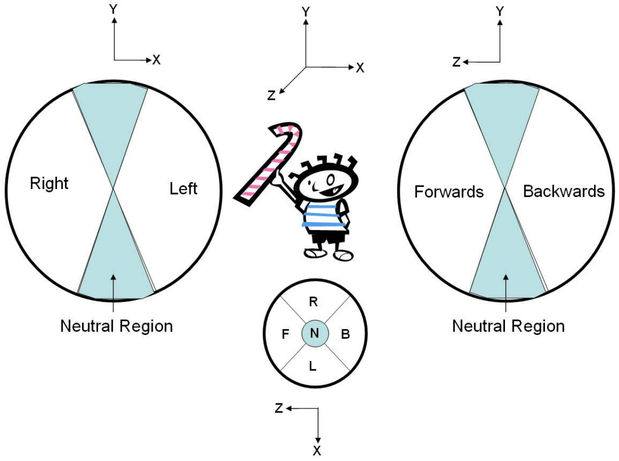

Figure 3: Angle regions

to the input integration module, which combines it with the speech input.

4.2 Speech Recognition

The speech recognition system uses the free speech recognition package Sphinx II from the Carnegie Mellon University. Sphinx is mature package, being in development for more than fifteen years, right now the most recent version is the Sphinx IV.

We chose the older Sphinx II, because it is fastest (the system is going to be used in the real time) and most robust out of the three versions (Sphinx II, III, IV), which are available.

Sphinx is a state–of–the–art large vocabulary, sub–word unit based, continuous speech recognizer, well suited e.g. for dictation or automated news transcription tasks. The complete description of the speech recognition engine can be found in [6, 13].

Our use of Sphinx is a bit non–typical, we trained the system to recognize just the keywords, not continuous speech. The reason is that our system is targeted to larger user base (e.g. an exhibition or public display, many users) and it is very difficult to obtain the necessary amount of training data for a speaker independent system. Moreover, there are many ways how people express the same thing, thus increasing the complexity of the problem even more.

Training just for keywords makes the system less intuitive to use, but greatly reduces the complexity of the problem and the amount of training data needed. Moreover, the magic spells are usually precisely formed words or expressions in the movies and tales we are trying to benefit from, so the impact on the intuitiveness of the use should not be large. But this has to be more precisely evaluated.

We use small vocabulary of approximately seventeen words and expressions, some examples are :

- “one”
- “two”
- “land”
- “fly”
- “left hand”
- “right leg”
- ...

Two acoustic models were built, one is English and the second one is French (our laboratory is in the french–speaking part of Switzerland), both languages can be used interchangeably, according to the preferences of the user.

The hardware used is a standard PC with a standard sound card, the speech is recorded via a headset worn by the user. We opted for a headset, which is much less convenient than a stand-alone microphone attached to the ceiling or fixed on a stand, because of the noise in the our laboratory, which could have an adverse effect on the recognition accuracy.

5 Results

We use the “magic wand” interface in a small adventure game, situated into the environment of the ancient Egypt, where the user is given the task of fulfilling some quests.

The virtual world is implemented using an integrated framework VHD++ [12], developed in the collaboration of VRlab, EPFL and MIRALab, University of Geneva.

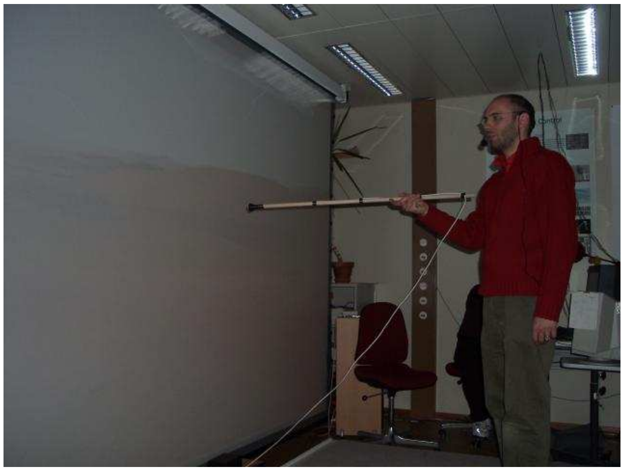

Figure 4: User flying around the landscape using the pointing mode

The game starts with the user flying a magic "carpet", exploring the landscape. This part uses the pointing mode of the wand in combination with the speech keywords "land" and "fly". The user navigates around the landscape by pointing at an object or in some direction and saying "fly". The software registers the direction the wand is pointing to, determines the object of interest the user is pointing at (if any – the user can point just to the ground or into the sky, in this case just the direction is used) and the flying "carpet" starts to move towards the designated target. In case, that the user desires to change the direction or discovers a new object of interest, he points the wand at the object and says "fly" again. The user does not control the carpet directly.

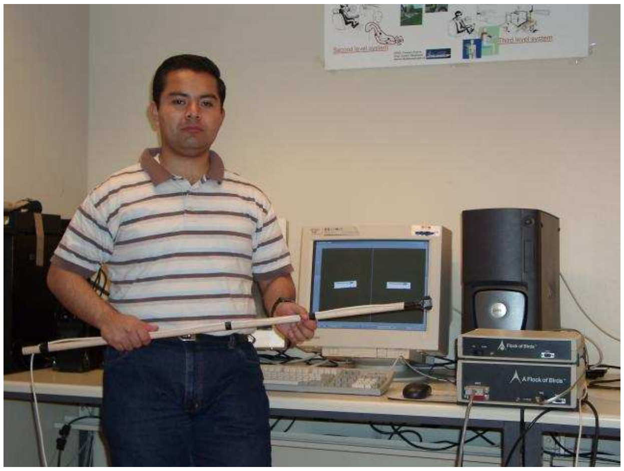

Figure 5: Monitoring application with FOB

The figure 4 shows the user flying outdoors in the virtual environment aboard the "carpet" (not visible, we just drive the camera directly for the moment).

The software driving the "carpet" is intelligent to some extent, it is able to avoid obstacles (e.g. the trees), keep the distance from the ground (not crash into the ground, hills or pass through the walls) and maintain the attitude to not confuse the user (e.g. by tilting the camera sideways).

It is very important for this navigation mode to be able to move the camera freely, since that is the only way the user is able to "discover" the points of interests . We didn't want to introduce a new device (e.g. joystick or a spaceball) just for that, because for a standing person in front of the projection screen it is very cumbersome, so we used a so–called "Go–Go technique", as described in [3]. While the wand is pointing inside of the large screen, the system just tracks the spot it is pointing at. When the wand moves outside of the screen, the system starts to move the camera in that direction (scrolling), the farther out of the screen the wand is pointing, the faster the scrolling, stopping immediately, when the wand points back to the screen.

The posture recognition mode is used indoors, letting the user select the direction he wants to follow to navigate inside of the virtual environment. The place is filled with different virtual characters demanding a variety of actions from the user in order to let him finish his quest. He has to solve puzzles and similar small games which require interaction with virtual objects: grabbing them, manipulating them (changing their position/orientation in space) etc.

The basic interaction mechanism consists of selecting the target object by voice and manipulating it with the wand. Only discrete actions (mainly translations) are required and for this the "magic wand" is the ideal interface (simple, and intuitive).

Our prototype shows some limitations of the technology used. For example, the speech recognition is not very accurate, with approximately 30 % error rate with the speaker–independent acoustic model. The most of problems are caused by noise in the laboratory, such as people chatting in the room or background music. Fortunately, this is not a critical problem, because the failures of the speech recognition do not cause errors in the interaction, the user just has to repeat the sentence again.

We observed, that the range of the magnetic sensor used is cca. 3 meters with the high power emitter, probably slightly more (we didn't test in larger space because of the space constraints of our laboratory). The position data returned from the magnetic sensor are very inaccurate, we are just using the orientation data. The accuracy of the sensor according to the manufacturer is $0.5°$, we didn't observe any problems related to tracking inaccuracies in our application. The system is sensitive to the environmental conditions in the place of installation, such as large metallic structures (steel reinforced concrete ceilings and floors or laboratory desks for example) and electromagnetic interference from various sources, which distort the magnetic field and cause errors. But these are usually static and can be compensated for.

A snapshot of the wand monitoring tool is shown in the figure 5 running in real time, with the current orientation of the wand on the left half of the screen and the recognized posture on the right half.

Figure 6 shows our prototype, with the large screen in the background, two out of three PCs used in the right part of the figure and the magic wand with the FOB emitter (the black cube) in the foreground.

Our application is still work in progress. The preliminary results are encouraging in comparison with the standard interfaces available (e.g. the 3D mice or joysticks).

6 Conclusions

Future work will include a public demonstration of the VR game described in section 5, in May 2003. This presentation will allow us to evaluate the "magic wand" as an intuitive interface for VR environments, because the visitors will be able to try and test the system and compare the ease of use with the direct control by a joystick, which will be available too.

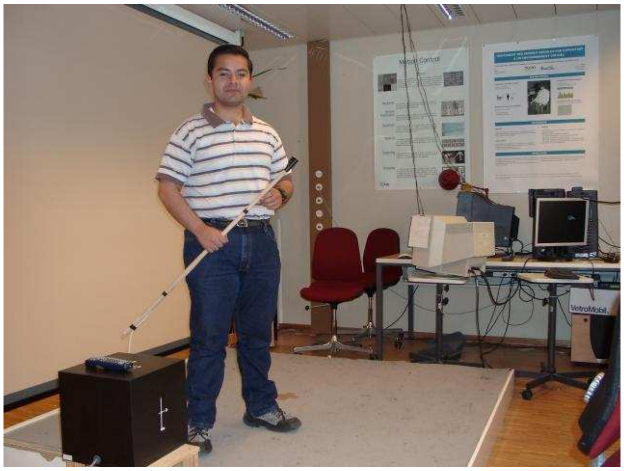

Figure 6: Laboratory setup

An improvement of the interface would be to substitute the use of magnetic sensors by optical tracking. This would increase the immersion on the virtual world (eliminating the wired sensor restricting the movements of the user).

Another possible improvement could be the on–line adaptation of the speech recognition system to the user, so that it "learns" from it's errors (unrecognized or incorrectly recognized words), thus improving the recognition rates. This could be implemented either by means of the user interrupting the work, switching to a "learning" mode and correcting the error of the system or by having the second user supervising the operations of the system and correcting the errors on–the–fly (e.g. from keyboard).

We have presented the preliminary results of a multimodal interface system for VR environments. Our main contribution is the implementation of a very simple and popular notion: using a "magic wand" in combination with "spell casting". This simple approach makes it easy for the regular user to start interacting in the virtual world, removing the learning curve of a less intuitive device such as a 3D mouse or a joystick. The "magic wand" interface has another advantage: it's a non–obstructive approach. The user doesn't need to wear any special equipment, "you just take your wand and cast the spell", neither it requires any special preparations of the user, e.g. training.

Finding the best way to interact in virtual environments is still an open issue. One of the main problems is the learning curve. Interfaces are usually not very intuitive, either because they don't resemble any device known by the regular user (3D mice or data–gloves are not common on every–day life), or because they restrict the user's motion and are uncomfortable to use (HMD, data-gloves). We believe a simpler, less obstructing approach that takes advantage of previous experience from the user, will lead to better results. The "magic wand" interface is trying to validate this idea.

7 Acknowledgments

Authors want to thank the team of designers of VRlab, especially to Rachel de Bondeli, Mireille Clavien and Stephanie Noverraz for designing the virtual environment, and to our colleagues Tolga Abaci for his work on event processing and Fatih Erol for the hard work on software integration.

References

[1] T. Auer, S. Brantner, and A. Pinz. The integration of optical and magnetic tracking for multi–user augmented reality. In Michael Gervautz, Dieter Schmalstieg, and Axel Hildebrand, editors, *Virtual Environments '99. Proceedings of the Eurographics Workshop in Vienna, Austria*, pages 43–52, 1999.

[2] R. A. Bolt. Voice and gesture at the graphic interface. In *ACM Computer Graphics 14,3*, pages 262–270, 1980.

[3] D. A. Bowman and L. F. Hodges. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *Symposium on Interactive 3D Graphics*, pages 35–38, 182, 1997.

[4] J. D. Brederson. *The I^3 Stick: An Inexpensive, Immersive, Interaction Device*. Center for Scientific Computing and Imaging, Department of Computer Science, University of Utah, Salt Lake City, UT84112-9205, 1999.

[5] A. D. Cheok et al. *Touch Space: Mixed Reality Game Space Based on Ubiquitous, Tangible, and Social Computing*. http://mixedreality.nus.edu.sg/research-EMRI-infor.htm.

[6] CMU Speech Group. *The CMU Sphinx Group Open Source Speech Recognition Engines*. http://www.speech.cs.cmu.edu/sphinx/.

[7] D. Krum, O. Ometoso, W. Ribarsky, T. Starner, and L. Hodges. Speech and gesture multimodal control of a whole earth 3d virtual environment, 2002.

[8] D. Levine, M. Facello, P. Hallstrom, G. Reeder, B. Walenz, and F. Stevens. Stalk: An interactive virtual molecular docking system, 1996.

[9] S. McGlashan and T. Axling. Talking to agents in virtual worlds. In *Proceedings of the 3rd UK VR-SIG Conference*, Leicester, England, 1996.

[10] A. Nijholt and J. Hulstijn. Multimodal interactions with agents in virtual worlds. In N. Kasabov, editor, *Future Directions for Intelligent Information Systems and Information Science, Studies in*

Fuzziness and Soft Computing, pages 148–173. Physica-Verlag, 2000.

[11] S. Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[12] M. Ponder, T. Molet, G. Papagiannakis, N. Magnenat-Thalmann, and D. Thalmann. VHD++ development framework: Towards extendible, component based VR/AR simulation engine featuring advanced virtual character technologies. In *Computer Graphics International 2003, to appear*.

[13] M. K. Ravishankar. *Efficient Algorithms For Speech Recognition*. PhD thesis, School of Computer Science, Computer Science Division, Carnegie Mellon University, Pittsburgh, PA 15213, 1996.

[14] Y. Sato, Y. Kobayashi, and H. Koike. Fast tracking of hands and fingertips in infrared images for augmented desk interface, 2000.

[15] J. Segen and S. Kumar. Human-computer interaction using gesture recognition and 3d hand tracking. In *International Conference on Image Processing*, Chicago, USA., 1998.

[16] E. Stephanie, S. Wauchope, and K. Perez. *A Natural Language Interface for Virtual Reality Systems*, 1996. http:// www.aic.nrl.navy.mil/ severett/.

[17] K. Wauchope. Spoken natural language agents for modsaf query and reference in quickset. Technical Report AIC-00-004, Naval Research Laboratory, Washington, DC, May 2000.