MULTI-MODAL NATURAL DIALOGUE

Proceedings of ACM CHI '90 Conference, Monterey, California, 1990 653-4.

Kristinn R. Thorisson David B. Koons Richard A. Bolt

The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, E15-404
Cambridge, MA 01239
kris@media-lab.media.mit.edu

INTRODUCTION

When people communicate with each other they use a wealth of interaction techniques. The multitudes of gestures, intonation, facial expressions, and gaze set the context for the spoken word, and are usually essential in human-to-human interaction [1, 5].

The Advanced Human Interface Group at the MIT Media Laboratory is exploring how the three modes of speech, gestures, and gaze can be combined at the interface to allow people to use their natural communication skills in interacting with the computer [2, 3]. The work is aimed at widening the means to communicate with computers and to make computing power available to the widest range of people.

The prototype system we have designed allows a person to interact in real-time with a graphics display by pointing, looking, asking questions, and issuing commands. The system responds with graphical manipulations and spoken answers to the user's requests.

INPUT TECHNOLOGIES

Speech

A head worn, noise-cancelling microphone feeds the speech to an AT 386 computer with hardware and software that allow for discrete word recognition. The recognized words are in turn sent to one of two host computers (Figure 2).

Hand

The VPL Dataglove™ gives information about finger posture of the user. The position and attitude of the hand is given by a magnetic sensor on the back of the hand. Pointing gestures are recognized on a host computer by a simple template-matching algorithm. A 3-dimensional vector extending out of the hand is intersected with the screen to find point of reference once a pointing gesture has been recognized.

Eye

To analyze the user's looking behavior we use a head mounted, corneal-reflection eye tracker. The user looks through a half-silvered mirror; an infrared LED light shines from above and lights up the eye. An infrared-sensitive camera picks up the reflection from the eye off the mirror and sends the resulting TV signal to an AT 286 computer for image processing. The resulting eye data is analyzed into *fixations*, *saccades*, and *blinks*.

Position of the head is found by using a magnetic sensor attached to the eye tracker. Fixations are sent to the host computer along with the user's head position during each fixation. These data are then combined to arrive at the user's point-of-gaze on a graphics screen (see Figure 1).

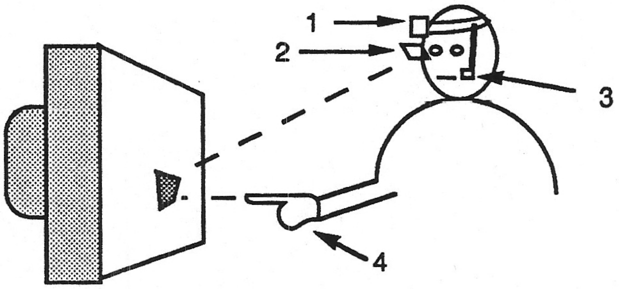

Figure 1.
An eye tracker (1=camera, 2=half-silvered mirror), microphone (3), and glove (4) allow the user to refer to objects on the screen by pointing, speaking, and looking.

SYSTEM DESCRIPTION

The system running on the two host computers has two basic components: an object-oriented map, with icons representing airplanes, trucks, helicopters, fire-fighting crews, and fire locations. The map is maintained by a graphics manager that keeps track of the position of all its icons, as well as their color, class, and name tag.

The second part of the system is a collection of modules called the *Agent*. The Agent can request information from the map manager about the layout of objects, and integrate it with the user's multi-modal requests. The Agent can thus arrive at an appropriate response for any request that the user issues. In the current version, the actions a user can perform are: *Move* an object to a new location, *delete* an object, *name* an object, *create* an object, and *request information* about the objects.

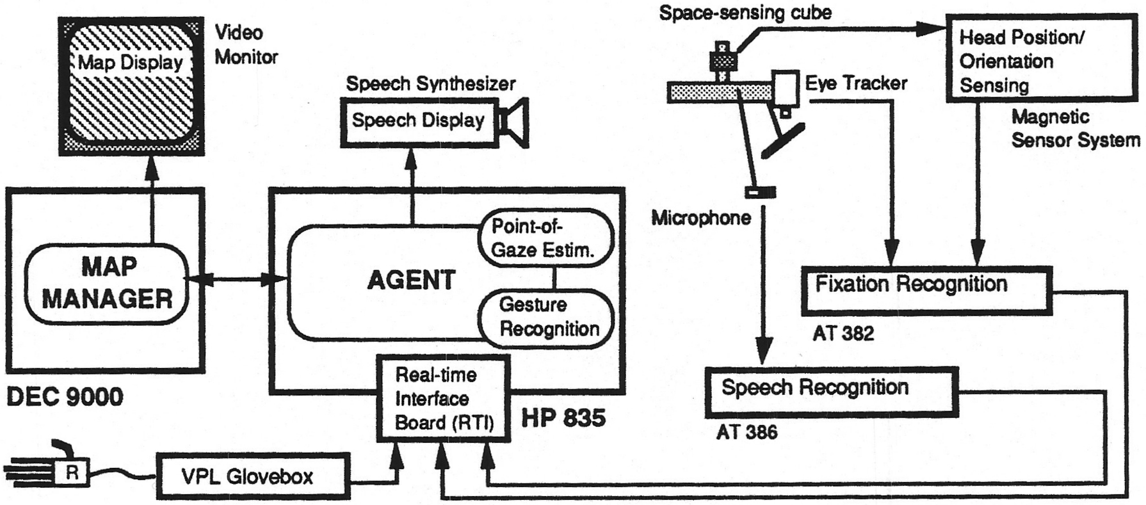

Figure 2.
System configuration, data links, and software modules.

Resolving Missing Information

Whenever the Agent receives a request that is under specified in speech, for example "delete *that* object," it will try to fill in the missing information by looking at what the user did in the two other modes around the time he said "that." The Agent looks for a pointing gesture and where on the screen the user fixated during that time. If one object is clearly singled out as the most likely referent, as indicated by the proximity of fixations and pointing to the object, then that object will be chosen as the referent and subsequently deleted.

When vital information is missing from the speech input, and the Agent cannot find a referent based on either hand or eye it will ask the user for further specifications.

Resolving Multiple Reference

The Agent can reason about relations between objects. This allows a user to say "delete the truck south of that fire." A reference is made to two objects, but one is derived by the location of the other.

"Move" commands involve two references; one to an object and one to a new location, as in "move *that helicopter* to *there*." The system can successfully deal with a continuous input of such commands by looking at the time that actions occurred and comparing it to the time that the user utters the important words of the phrase (in this case "that helicopter" and "there").

FUTURE DIRECTIONS

By allowing for multi-modal interaction, people can use their social skills in interacting with the computer. We will be looking at further ways to make such interaction possible; giving the computer a greater sense of two-handed gestures as they occur naturally in 3-dimensional space, and the role of gaze in communication. Other tasks include giving the Agent a memory, a greater understanding of spatial relationships, and a face that can glance back at the user [4].

ACKNOWLEDGEMENTS

The authors acknowledge the contributions of graduate students Brent C.J. Britton and Edward Herranz, and assistants David L. Berger, Michael P. Johnson, Mathew Kaminski, Brian Lawrence, and Christopher Wren, and research affiliate Masaru Sugai, NEC, Japan.
This research was supported by the Defence Advanced Research Projects Agency under Rome Laboratories, contract F30602-89-C-0022.

REFERENCES

1. Argyle, M. & Cook, M. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, England, 1975.

2. Bolt, R. A. *The Human Interface*. Lifetime Learning Publications, Belmont, CA, 1984.

3. Bolt, R. A. The Integrated Multi-Modal Interface. In *Transactions of the Institute of Electronics, Information, and Communication Engineers* (Japan), (Nov. Vol. J70-D, No. 11, 1987), pp. 2017-2025.

4. Britton, Brent C.J. *Enhancing Computer-Human Interaction With Animated Facial Expressions*. Unpublished Master's Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1991.

5. Nespoulous, J-L, & Lecours, A. R. Gestures: Nature and Function. In *The Biological foundations of Gestures: Motor and Semiotic Aspects*, J-L. Nespoulous, P. Perron, & A. Roch (eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 1986, pp. 49-62.