Multimodal User Interface for Mission Planning

A. Medl, I. Marsic, M. Andre, C. Kulikowski and J. Flanagan
Rutgers University, CAIP Center
96 Frelinghuysen Rd.
Piscataway, NJ 08854-8088
medl@caip.rutgers.edu

Abstract

This paper[1] presents a multimodal interface featuring fusion of multiple modalities for natural human-computer interaction. The architecture of the interface and the methods applied are described, and the results of the real-time multimodal fusion are analyzed. The research in progress concerning a mission planning scenario is discussed and other possible future directions are also presented.

Introduction

Current human-machine communication systems predominantly use keyboard and mouse inputs that inadequately approximate human abilities for communication. More natural communication technologies such as speech, sight and touch, are capable of freeing computer users from the constraints of keyboard and mouse. Although they are not sufficiently advanced to be used individually for robust human-machine communication, they have adequately advanced to serve simultaneous multisensory information exchange (Cohen et al. 1996, Waibel et al. 1995). The challenge is to properly combine these technologies to emulate the natural style of human/human communication by making the combination robust and intelligent (Flanagan and Marsic 1997).

Problem Statement

The objective of this research is to establish, quantify, and evaluate techniques for designing synergistic combinations of human-machine communication modalities in the dimensions of sight, sound and touch in collaborative multiuser environments (see Fig. 1).

The CAIP Center's goal is to design a multimodal human-computer interaction system with the following characteristics and components:

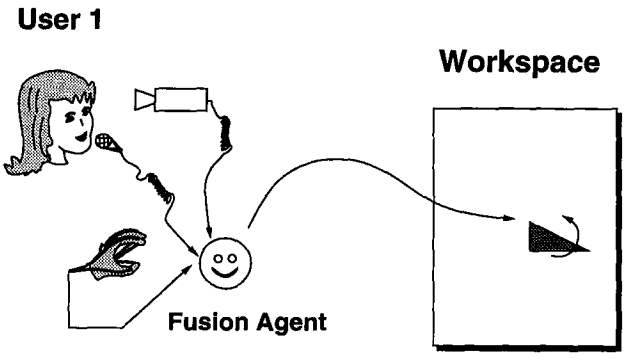

Figure 1: Multimodal interaction.

- force-feedback tactile input and gesture recognition
- automatic speech recognition and text-to-speech conversion
- gaze tracking
- language understanding and fusion of multimodal information
- intelligent agents for conversational interaction and feedback
- applications for collaborative mission planning and design problems

Although the system is in an early stage of development, results concerning tactile information processing, robust gesture recognition, natural language understanding, gaze tracking, and multimodal fusion are promising.

The present study combines real-time speech input with gaze input, as well as asynchronous gesture input provided by the CAIP Center's Rutgers-Master II force-feedback tactile glove (Burdea 1996, Gomez et al. 1995).

Real-life problems of collaborative mission planning are being tested as potential applications of the multimodal interface. Ongoing research concerning mul-

[1]Copyright 1997, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

timodal manipulation of icons on military maps is reported.

First, we briefly describe the functional system components. Then, the methods applied for language processing and sensory fusion are introduced and research progress is outlined.

System Components

RM–II Force-Feedback Tactile Glove

The Rutgers Master II (RM–II) system is a portable haptic interface designed for interaction with virtual environments (Burdea 1996). Its two main subsystems, shown in Fig. 2, are the *hand–master* and the *Smart Controller Interface* (SCI).

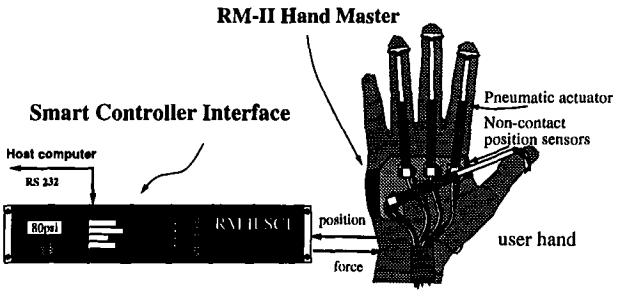

Figure 2: The Rutgers Master-II force-feedback tactile glove.

The RM–II hand master can read hand gestures (fingertip positions relative to the palm) and apply forces on the fingertips corresponding to the interaction.

The hand–master unifies hand motion sensing and force display in a small, compact structure weighing 80 grams. The main structure consists of a small "L" shaped palm base on which four custom-designed linear-displacement pneumatic actuators and 8 Hall-effect sensors are mounted. The actuators are equipped with a special sensor to measure their linear displacement. A 6D Polhemus tracker (Polhemus 1993) mounted on the back of the hand provides wrist position/orientation.

Hand gesture module

The hand gesture module is implemented using the RM–II system described above. Hand gesture input is particularly efficient for two kinds of tasks: 3D environment surfing (changing the viewpoint of virtual environment) and virtual objects manipulation. Since the glove was used in this application in a 2D graphic environment, only gestures for object manipulation were implemented. In addition, a special gesture was defined to emulate mouse clicks (to maintain compatibility with current GUI's). The design of the hand gesture module features only natural hand gestures in an effort to create an easy-to-use interface (Pavlovic et al. 1997).

The first step in designing the hand gesture module involves implementing an object *selection function.* In a 2D environment, the coordinates are calculated where a virtual ray along the user's index finger intersects the screen plane and outputed when the corresponding gesture ("pointing") is executed. The selection point is further checked to see if it is inside any object in the environment. Other gestures designed for object manipulation are:

grab : used for grabbing and moving the selected object;

thumb up : associated with resizing the selected object;

open hand : corresponds to the 'unselect' or 'drop object' command; it is also used as a reset position for hand gestures;

curling the thumb : corresponds to mouse clicks.

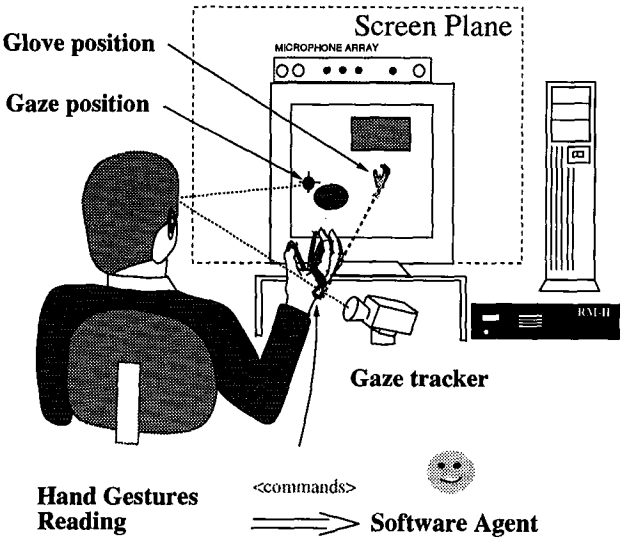

Figure 3: Hand gesture interaction and glove display.

Once a hand gesture is recognized, the corresponding string is generated and sent to the parser as indicated in Fig 4. In addition, a separate stream is continuously sending hand-pointing screen coordinates.

Gaze tracker

The gimbal mounted gaze tracker (ISCAN 1994) consists of a camera and an infrared illuminator. The illuminator projects infrared light into the left eye of

the user while a sophisticated image processing software continuously analyzes the video image and tracks the gaze by computing the angle between the centroid of the pupil and the corneal reflection.

After proper calibration the gaze tracker outputs x-y screen coordinates. These coordinates are forwarded to the fusion agent for further analysis and to the application, where an eye-cursor is continuously displayed on the screen.

Speech Recognizer

The current system uses a Microsoft speech recognizer engine (Microsoft 1996) with a finite-state grammar and a restricted task-specific vocabulary. The recognition is speaker-independent.

Once the speech recognizer has recognized a phrase, it sends the text to the parser together with two time-stamps. Although the current recognizer does not provide time-stamps after each word, it does provide the time at the start and end of each utterance. In future applications – to achieve temporal synchronization of the modalities –, time-stamps after every word of the utterance will be necessary to exactly determine where the user is pointing at (with tactile glove or gaze) while speaking.

Furthermore, the Whisper system exclusively runs under Microsoft Windows and is not portable to different platforms. Therefore, a CAIP-developed recognizer (Lin 1996) will be applied to solve these problems.

Microphone Array

CAIP's microphone array technology liberates the user from body–worn or hand–held microphone equipment, permitting freedom of movement in the workplace. The current fixed focus line microphone array focuses on the speaker's head sitting approximately 3 feet from the monitor. Other sound sources are successfully attenuated. A CAIP-developed microphone array system is applied as a robust front-end for the speech recognizer to allow distant talking (Lin 1996).

Language Processing and Sensory Fusion

Parser

The first step in the understanding of multimodal commands involves parsing of the sensory inputs. In our system, the parser communicates with each modality and the fusion agent as illustrated in Fig 4. The reason for communicating through the parser is that we have chosen spoken language as the common means of communication. Gesture information provided by the tactile glove is first translated into written text by the gesture recognition module, and forwarded for further analysis to the parser. Note that this process is similar to the translation of sign-language gestures into their spoken language representations.

In the test application, the initial task-specific vocabulary is small and contains about 150 words. We implemented a simple context-free grammar as shown below. The following code fragment uses a common syntax in which:

- capitalized identifiers are non-terminals
- parenthesized identifiers are terminal symbols
- alternatives are separated by the symbol |
- a non-terminal between rectangular brackets [ ] is optional
- PHRASE is the start symbol

```
PHRASE
= [ REDUNDANT ] META_PHRASE [ REDUNDANT ]
| [ REDUNDANT ] CMD_PHRASE [ REDUNDANT ]
| NOUN_PHRASE
| PREP_PHRASE
| REDUNDANT

REDUNDANT
= "can you" | "please" | "I want to"
| "OK" | "with" | "of"

META_PHRASE
= "no" | "undo" | "undo last command"
| "forget it" | "leave it" | "end"

CMD_PHRASE
= CMD_VERB [ NOUN_PHRASE ] [ PREP_PHRASE ]
[ PREP_PHRASE ]

CMD_VERB = "move"  | "relocate"
| "remove" | "delete"
| "resize"
| "draw"  | "create"
| "place" | "put"

NOUN_PHRASE
= [ ARTICLE ] [ MODIFIER ] OBJ_TYPE
| POSITION_SPEC

ARTICLE = "a" | "an" | "the"

MODIFIER = COLOR [ SIZE ]
| SIZE [ COLOR ]
| "new"

COLOR = "red" | "green" | "blue" | "black"
```

```
SIZE = "short" | "small"
| "big" | "large" | "long"


OBJ_TYPE = "rectangle" | "box"
| "circle" | "oval"
| "line"

PREP_PHRASE = PREPOSITION POSITION_SPEC

PREPOSITION
= "to" | "from" | "at" | "through"
| "passing through" | "till"
| "up till" | "up to"

POSITION_SPEC   = "here" | "there" | "this
position" | "this point"
| TERRAIN_FEATURE

TERRAIN_FEATURE = "hill" ID | "camp" ID

MODALITY        = "with gaze" | "with glove"
```

As mentioned above, synchronization of speech and gesture is essential in future developments. Currently, glove positions needed for the application are indicated by pointing gestures of the tactile glove and forwarded to the parser as position specifiers. Also, the position of the eye-cursor is forwarded to the parser.

Multimodal Fusion

Our approach to multimodal fusion is a variation of the slot-filler method which is well-known in artificial intelligence (Winston 1992, Allen 1995).

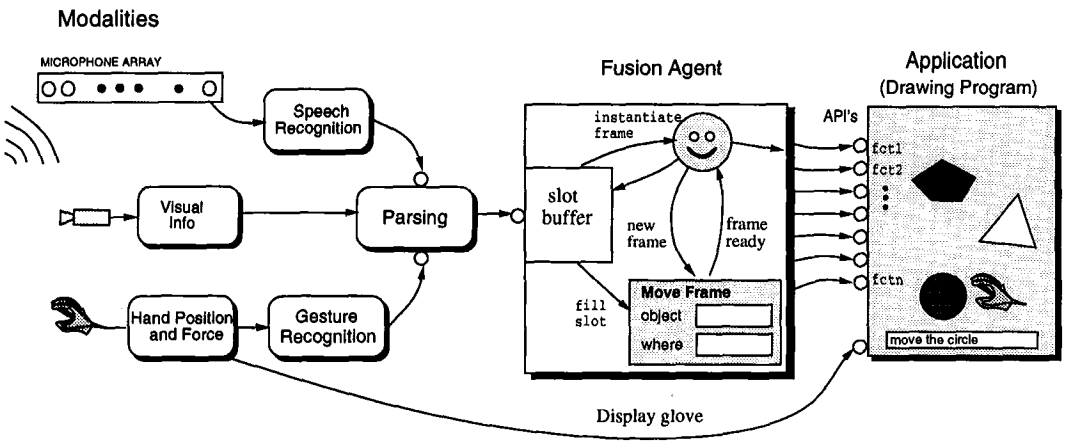

Figure 4: Integration of tactile glove, speech, and gaze input in a collaborative application.

Information necessary for command or language understanding is often encoded in structures called *frames*. In its most abstract formulation, a frame is simply a cluster of facts and objects that describe some typical object or situation. In our application, frames contain information about commands given by multiple modalities, such as tactile glove and speech inputs. A similar architecture is described in (Shaikh 1996).

The principal objects in a frame are assigned names, called *slots*. For example, the command frame "delete" contains only one slot called "object ID," while the frame "create" should contain slots called "object type," "color," etc. The slots can be viewed as functions that take values to instantiate the actual frame. Thus a particular instance of a frame represents a structured set of knowledge.

Our approach to multimodal fusion is illustrated in Fig. 4. We are using a predefined set of frames corresponding to the possible commands in the current application. A slot writer-reader method which installs and retrieves slot values has been implemented as part of the fusion agent module. It uses the keyword approach combined with some grammar constraints.

Slot Buffer. An important component of the fusion agent is the slot buffer. It stores the incoming values for all possible slots defined by the command vocabulary. This is essential because present information is often used in the future. Consider this example: the user selects an object by saying "Select the green circle." Next, he can say "Move it." If the object from the previous utterance is stored in a slot buffer, then the "Move it" command has sufficient information about what to move. In other words, the slot buffer contains memory of past operations.

Slot Filling and Command Execution. The parser fills the slots in the slot buffer that are des-

ignated in the utterance using the slot-writer method and reads the tactile glove positions and gestures when appropriate.

For example, the utterance "*From here to here create a red rectangle,*" causes the following slots to be filled in the slot-buffer: the positions of the two opposite corners of the object, the object's type, the object's color, and the operation or command type.

On the other hand, if only the gesture recognition module is providing information, the parser relies only on that input. In the current design, complementarity and concurrency problems are handled on a first-come first-serve basis. The system executes a command as soon as all necessary information is available. The information may be provided entirely or partly by either modality.

The instantiation of a particular command frame is done by analysis of information in the slot buffer. A demon monitors the slot buffer to determine if the command slot is filled. If it is not filled, the system will wait for more input information provided by any modality.

If the command slot is filled, the demon will instantiate the corresponding command frame and examine if there is sufficient information in the slot-buffer to fill each predefined slot of that particular frame. The demon must of course wait until all slots are filled. Then the command will be executed through the application programming interface (API).

To achieve maximum flexibility, the design assures that a particular frame contains the minimum number of slots that are necessary to unambiguously execute a command. For example, if there is a *blue rectangle* and a *red circle* in the workspace, it is sufficient to say "*delete the rectangle*" because the "*delete*" frame only contains the minimum necessary information i.e., the "*object ID*" slot. The utterance "*delete the large blue rectangle*" generates the same command because it obviously contains the minimum information needed for the intended action.

If the command slot of the slot buffer is not filled by the parser, then the system will wait for more input information provided by either modality.

Results and Discussion

Our current testbed uses a 150-word vocabulary with a finite grammar and a force-feedback tactile glove. The speech recognition works adequately for distant talking due to the microphone array. The fusion agent with the slot-buffer performs well even for asynchronous inputs from modalities. The command execution in the collaborative drawing program operates without substantial delays when tested in our laboratory. However, its current functionality is limited to manipulating (creating, moving, resizing, deleting, etc.) colored geometric objects and icons on topographical maps.

Fig. 5 illustrates how colored objects can be manipulated using simultaneous speech, gaze, and tactile input. Figs. 6 and 7 show icons being moved by a command generated by multiple modalities. Here, "helicopter 96" is selected and moved by the "*grab-move*" gesture, gaze, and speech.

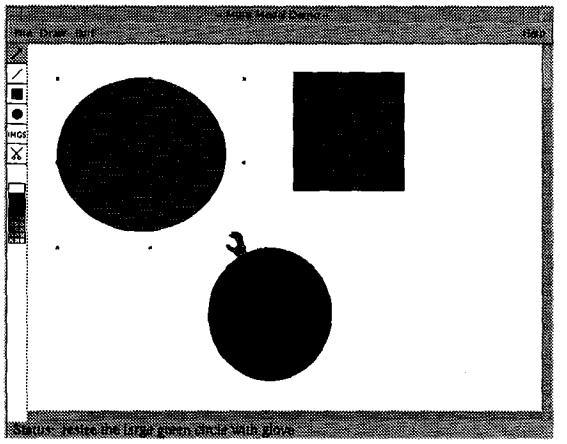

Figure 5: Manipulation of colored objects by speech and gesture in a collaborative drawing program.

Research in Progress

We intend to integrate a gaze-tracker to select and manipulate objects, adding further flexibility to the sensory input.

We are also proposing to investigate the possibility of a gesture recognition agent design to interpret possible commands given by gaze sequences.

A three-dimensional graphical environment will be implemented to simulate 3D problems, tissue palpation, and other medical applications.

An intelligent information agent is being designed to answer queries and provide information to the user about the actual state of the workspace (e.g., contents, positions and types of icons, history of past operations, etc.). This will be a crucial element of a mission planning system which could be applied to disaster relief or rescue operations.

Text-to-synthetic speech answerback to the user will include notification about unrecognized phrases and unexecutable operations as well as acknowledgment of commands.

An important aspect of networked multimodal systems is collaborative work. The DARPA-sponsored effort *DI*stributed *S*ystem for *C*ollaborative *I*nformation *P*rocessing and *LE*arning (DISCIPLE) has already resulted in a framework for multiuser collaboration. We

are planning to apply the methods of multimodal human-computer interaction in collaborative and cooperative work in the near future.

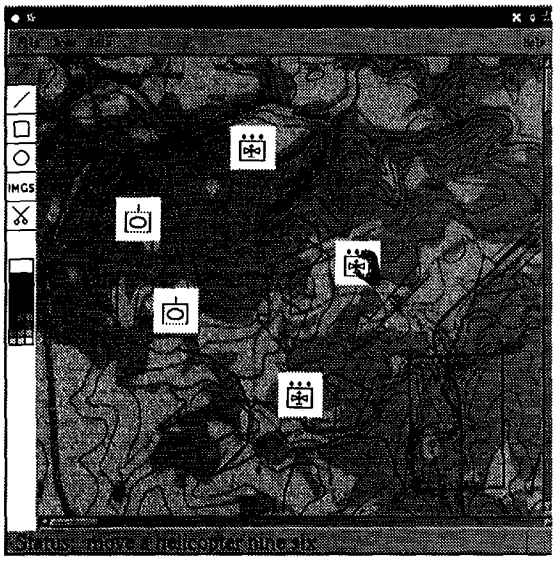

Figure 6: Manipulation of military icons by speech and gesture in a collaborative environment.

A mission planning scenario is being designed to evaluate the multimodal collaborative system. Geographical information is represented in map-tile format similar to that of the MARCO system (Samet and Soffer 1996). In order to achieve interactivity, the user should be able to indicate self-defined regions with the tactile glove pointer and speech. For example, the utterance "now I am indicating camp number six" results in activating a recording program which fills in appropriate map-tiles with information while the user simultaneously leads the pointer along the region to be defined as "camp number six." The utterance "finish recording" deactivates the recording program. The information is stored in memory, so the user can refer to that region (e.g., "move tank four six to camp number six").

A structure for knowledge representation is being designed to include a description of **a)** the map-content (i.e., topographical or man-made features) which is expected to remain invariant during a problem-solving session, **b)** the icons expected within the session, including their attributes and hierarchical decomposition into subicons, **c)** rules constraining the expected and allowed behavior of icons within the map-context.

Various architectures for the intelligent fusion agent (probabilistic decision networks, adaptive classification and neural networks) are being considered to achieve robust and intelligent command interpretation and feedback to the user.

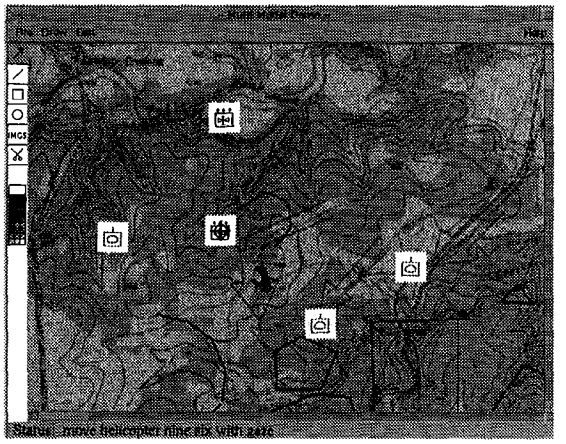

Figure 7: Manipulation of military icons by speech and gaze in a collaborative environment.

Acknowledgment

The authors had helpful discussions with Professors G. Burdea and J. Wilder. Special thanks to G. Popescu, A. Shaikh, and Y. Liang for their assistance in system implementation.

Components of this research are supported by NSF Contract No.IRI-9618854, DARPA Contract No.N66001-96-C-8510, and by the Rutgers Center for Computer Aids for Industrial Productivity (CAIP). CAIP is supported by the Center's corporate members and the New Jersey Commission on Science and Technology.

References

Books

Allen, J. 1995. *Natural Language Understanding*, California: The Benjamin/Cummings Publ. Co.

Burdea, G. 1996. *Force and Touch Feedback for Virtual Reality*, New York: John Wiley & Sons.

Winston, P.H. 1992. *Artificial Intelligence*, 3rd edition, Addison-Wesley Publ. Co.

Journal Article

Pavlovic, V.I.; Sharma, R.; and Huang, T.S. 1997. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7).

Samet, H.; and Soffer, A. 1996. MARCO: MAp Retrieval by COntent. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(8):783-798.

Waibel, A.; Vo, M.T.; Duchnowski, P; and Manke, S. 1995. Multimodal Interfaces. *Artificial Intelligence Review*, 10(3-4).

Proceedings Papers

Cohen, P.R.; Chen, L.; Clow, J.; Johnston, M.; McGee, D.; Pittman, J.; and Smith, I. 1996. Quickset: A Multimodal Interface for Distributed Interactive Simulation. In Proceedings of the UIST'96 demonstration session, Seattle.

Flanagan J.L.; and Marsic, I. 1997. Issues in Measuring the Benefits of Multimodal Interfaces. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97), Munich, Germany.

Gomez, D.; Burdea, G.; and Langrana, N. 1995. Modeling of the RUTGERS MASTER–II haptic display. In Proceedings of ASME WAM, DSC-Vol. 57-2: 727–734.

Lin, Q.; Che, C-W.; Yuk, D-S.; and Flanagan, J.L. 1996. Robust Distant Talking Speech Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), Atlanta, GA, pp.21-24, May 1996.

Shaikh, A.; Juth, S.; Medl, A.; Marsic, I.; Kulikowski C.; and Flanagan, J.L. 1997. An Architecture for Fusion of Multimodal Information. In Proc. Workshop on Perceptual User Interfaces, Banff, Alberta, Canada.

Technical Reports

ISCAN, Inc. 1994. Operating Instructions.

Polhemus, Inc. 1993. Fastrak User's Manual. Colchester, VT.