

An Architecture for Multimodal Information Fusion

A. Shaikh, S. Juth, A. Medl, I. Marsic, C. Kulikowski, and J. L. Flanagan
CAIP Center, Rutgers University
96 Frelinghuysen Road, Piscataway, NJ 08854-8088
medl@caip.rutgers.edu

Abstract

This paper 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.

1 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 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 [2], [6]. The challenge is to properly combine these technologies to replicate the natural style of human/human communication by making the combination robust and intelligent [3].

2 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.

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

- 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 and multimodal fusion are promising.

The present study combines real-time speech input with asynchronous gesture input provided by the CAIP Center's *Rutgers-Master II* force-feedback tactile glove [1]. The system will be extended with a gaze tracker in the near future.

3 System Components

3.1 RM-II Force-Feedback Tactile Glove

The Rutgers Master II (RM-II) system is a portable haptic interface designed for interaction with virtual environments [1]. Its two main subsystems, shown in Fig. 1, are the *hand-master* and the *Smart Controller Interface* (SCI).

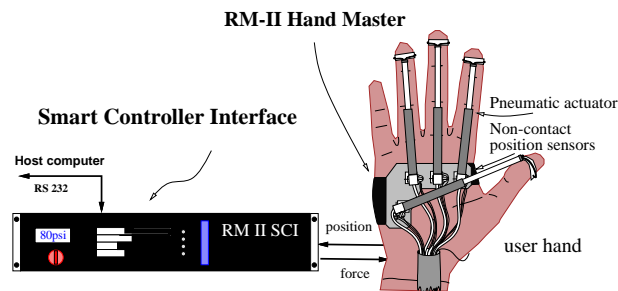


Figure 1: The Rutgers Master-II force-feedback tactile glove.

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.

Hand gesture module

The hand gesture module is implemented using the RM-II system described above. 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 outputted 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.

Once a hand gesture is recognized, the corresponding string is generated and sent to the parser as indicated in Fig 3. In addition, a separate stream is continuously sending hand-pointing screen coordinates.

Speech Recognizer

The current system uses a Microsoft speech recognizer engine [5] 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 mouse) while speaking.

Furthermore, the Whisper system exclusively runs under Microsoft Windows and is not portable to different platforms. Therefore, a CAIP-developed recognizer [4] 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 [4].

3.2 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 3. 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.

Multimodal Fusion

Our approach to multimodal fusion is a variation of the slot-filler approach where a slot-buffer stores the incoming values for all possible slots defined by the command vocabulary. First the parser fills the slots that are designated in the utterance and reads the mouse position 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.

A demon is watching the slot-buffer to see if the command slot is filled. If it is filled, then it will instantiate the appropriate command frame and examine if there is enough information in the slot-buffer to fill the predefined slots of that particular frame. Then the command will be executed through the application interface. If it is not filled, then the system will wait for more information. The block diagram of the current architecture is illustrated in Fig. 3.

4 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. 2 shows an icon of a helicopter being moved by a command generated by the gesture recognition module and the speech recognizer. Here, "helicopter 96" is selected by speech and moved by the "*grab-move*" gesture of the tactile glove.

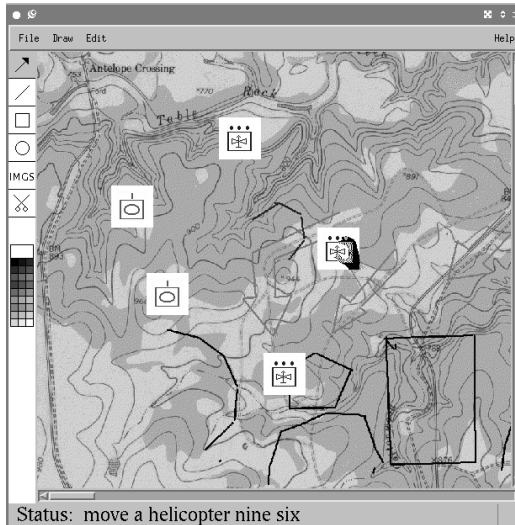


Figure 2: Manipulation of military icons by speech and gesture in a collaborative environment.

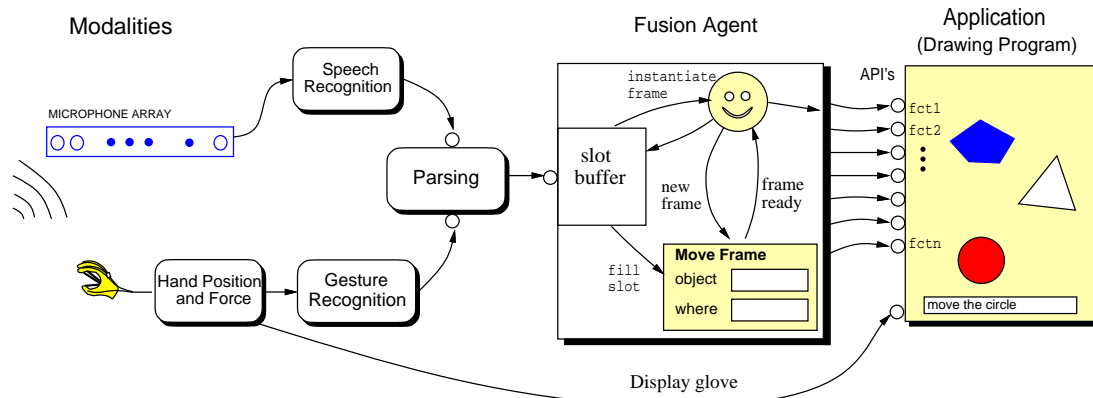


Figure 3: Integration of tactile glove and speech input in a collaborative application.

5 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.

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 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. For more information, see:
<http://www.caip.rutgers.edu/multimedia/multimodal/>

6 Acknowledgment

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).

References

- [1] G. Burdea, *Force and Touch Feedback for Virtual Reality*, John Wiley & Sons, New York, 1996.
- [2] P. R. Cohen, L. Chen, J. Clow, M. Johnston, D. McGee, J. Pittman, and I. Smith, "Quickset: A Multimodal Interface for Distributed Interactive Simulation," *Proceedings of the UIST'96 demonstration session*, Seattle, 1996.

- [3] J. L. Flanagan and I. Marsic, "Issues in Measuring the Benefits of Multimodal Interfaces," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Munich, Germany, April 1997.
- [4] Q. Lin, C.-W. Che, D.-S. Yuk, and J. L. Flanagan, "Robust Distant Talking Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta, GA, pp.21-24, May 1996.
- [5] Whisper Speech Recognizer by Microsoft Corp., www.research.microsoft.com/research/srg/whisper.htm
- [6] A. Waibel, M. T. Vo, P. Duchnowski, and S. Manke, "Multimodal Interfaces," *Artificial Intelligence Review*, Vol.10, No.3-4, 1995.