An Architecture for Fusion of Multimodal Information


Correspondence:

Dr. Attila Medl
Rutgers University, Center for Computer Aids for Industrial Productivity (CAIP)
Frelinghuysen Rd., P.O.Box 1390, Piscataway, NJ 08855-1390, USA
Tel: (908)-445-0080, Fax: (908)-445-4775
E-mail: medl@caip.rutgers.edu
An Architecture for Multimodal Information Fusion\(^1\)

A. Shaikh, S. Juth, A. Medli, I. Marsic, C. Kulikowski, and J. L. Flanagan
CAIP Center, Rutgers University, Piscataway, NJ

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], [4]. 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]. This study combines real-time speech input with asynchronous gesture input provided by the mouse pointer. The use of mouse buttons has been minimized to ease its functional replacement with the CAIP Center’s Rutgers-Master II force-feedback tactile glove [1], and a gaze tracker in the near future.

2 System Architecture

2.1 Speech Communication Interface

The current system uses the Microsoft speech recognizer engine with a finite-state grammar and a restricted problem-specific vocabulary. A CAIP-developed microphone array system is applied as a robust front-end for the speech recognizer to allow distant talking.

2.2 Language Processing and Sensory 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 Figure 1.

3 Results and Discussion

Our current testbed uses a 150-word vocabulary with a finite grammar and a mouse pointer. The speech recognition works adequately for distant talking due to the microphone array.

\(^1\)The research reported here is supported by the NSF Contract No. IRI-9618854 and by the Center for Computer Aids for Industrial Productivity (CAIP). The CAIP Center is supported by the New Jersey Commission on Science and Technology and the Center’s Industrial Members.
The fusion agent with the slot-buffer works well even for asynchronous inputs. 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 geometric objects (creating, moving, resizing, deleting, etc.) and is being extended to medical image processing.

The text-to-speech feedback to the user includes notification about unrecognized phrases and unexecutable operations as well as acknowledgement of commands.

4 Research in Progress

Integration of the unique force-feedback tactile glove Rutgers-Master II [1] as a modality is a key part of the ongoing research. A gesture recognition agent is being designed to interpret hand-gestures and fill the appropriate slots in the slot-buffer.

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

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

New 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.

References


