A Multimodal Human-Computer Interface for the Control of a Virtual Environment

Gregory A. Berry, Vladimir I. Pavlović[1], and Thomas S. Huang
Image Formation and Processing Group
Beckman Institute and ECE Department
University of Illinois at Urbana-Champaign
405 N. Mathews Avenue, Urbana, IL 61801
{berry,vladimir,huang}@ifp.uiuc.edu
[1]**Corresponding Author**: tel. (217) 244-1089, fax (217) 244-8371

Abstract

To further the advances in Human Computer Intelligent Interaction (HCII), we employ an approach to integrate two modes of human-computer communication to control a virtual environment. By using auditory and visual modes in the form of speech and gesture recognition, we outline the control of a task specific virtual environment without the need for traditional large scale virtual reality (VR) interfaces such a wand, mouse, or keyboard. By using features from both speech and gesture, a unique interface is created where different modalities complements each other in a more "human" communication style.

Introduction

Since the advent of the computer, the user has been forced to conform to the interface dictated by the machine. For the ancient Chinese the interface was the beads of the abacus, and today the modern interfaces interact with the user through a mouse movement, key press, or even speech input. These interfaces can be efficient for a trained user, but they are typically inefficient as a human-centric form of communication. For example when communicating with a friend, we would rather see the person and converse with them, rather than staring at a terminal full of characters while typing a message.

With recent advances in multi-modal interaction it has become more feasible to create interfaces that resemble forms of human communication. Although it is still impossible to create an ubiquitous interface that can handle all forms of human communication, it is possible to create a small multi-modal subset. Most modern interfaces rely exclusively on a single mode of interaction such as speech or mouse. We propose a more "natural" environment where speech in conjunction with free hand gestures provides the interface. Several systems have recently emerged that integrate multiple interface modalities in desktop/hand-held-based computer environments (Cohen *et al.* 1997). However, a perfect application for such an interface is in the use of large scale immersive displays or virtual reality (VR) environments such as the CAVE and Immersadesk technologies. Psychological studies in virtual environments have shown that people prefer to use hand gestures in combination with speech (Hauptmann & McAvinney 1993). Currently in these immersive applications, 3-D motion sensors and a wand form the user interface to an immersive display. We believe a more natural interface is created by replacing the wand with a set verbal and gestural commands. By removing the wand with it's mechanical cords and buttons, we are making the interface less encumbered and easier to use.

System Integration

As stated previously, creating a completely generic human-computer interface (HCI) is at the least a daunting, if not impossible task. To make this task manageable, rather than define a generic interface, we constrain the interface by the task. Therefore, we limit the commands, auditory and visual, to the specific domain of the task. As a testbed for our implementation we have chosen to implement an interface for a VR environment called BattleView (see Figure 2.) BattleView was created by the National Center for Supercomputing Applications (NCSA) for studying graphic display and user interaction strategies in support of planning and decision-making tasks in a virtual battlefield (Baker).

BattleView restricts the interface to a simple command domain that requires user movement and object selection in a 3D space. To model this interface consider an auditory/visual HCII system, symbolically depicted in Figure 1. The system consists of three modules: visual, auditory, and the multimodal integration module. The auditory module is designated for automatic speech recognition (ASR), while the visual module interprets moving actions of the human arm and hand through automatic gesture recognition (AGR). Integration of the two modes defines the multimodal integration module.

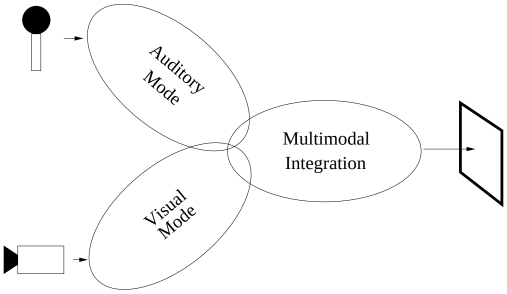

Figure 1: Integrated multimodal interface structure. Variable cardinality represents different command structures

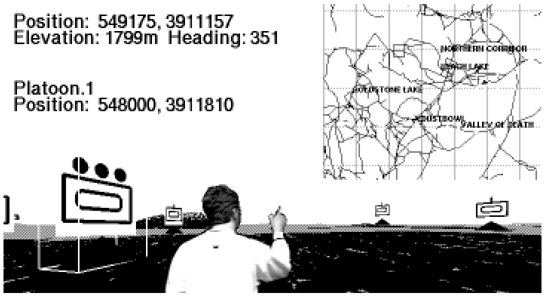
Position: 549175, 3911157
Elevation: 1799m Heading: 351
Platoon.1
Position: 548000, 3911810

Figure 2: BattleView virtual simulation environment.

Visual Module

The visual module for AGR is designed to recognize two forms of gesture. The first is simple pointing. By using a fixed point and location of the hand with respect to the head, a 3D ray is created. The intersection of the ray with the display creates a pointer that can be used for object selection. The hand is essentially a 3D mouse. The second form of gesture recognized is the more formal recognition of gesture, where defined hand movements are converted to commands via a visual feature classifier. The visual feature classifier dynamically quantizes temporal patterns hand movements according to some prior knowledge. Hidden Markov Models (HMM's) (Rabiner 1989) are used for this task as their special case provide invaluable modeling tools for scale and shift invariant temporal pattern classification. To achieve robust, real time feature tracking for the hand in both forms of gesture, we used skin color region segmentation based on color predicates (Kjeldsen & Kender 1996) and a Kalman filter based second order tracking algorithm as described in (Pavlović, Berry, & Huang 1997).

Auditory Module

For the auditory module, we selected a typical feature estimator/classifier architecture for speech recognition systems, similar to the one in (Sharma *et al.* 1996). A simple command grammar was defined, and word level PIN's where again selected as the feature classifier. To facilitate HMM development, a commercially available *HTK Toolkit* from Entropic Research was used for our "classifier module" design.

Integration Module

Integration of gestures and speech is carried out in the multimodal integration module. In this module, input states normally reserved for specific buttons, joystick, and other wand type functionalities, are tracked and manipulated by audio and gestural commands. For instance, to start movement in the virtual environment for flying, the user utters "fly". When this verbal command is recognized, the gesture recognizer changes its state from a feature classifier to 3D point mode. Now the gesture module input is being used to navigate as a plane's joystick. By the position of the hand with reference to the head, the gesture now defines an airplane-like control. To change the speed and direction, the user utters "forward, faster". By using a command grammar that incorporates each of the modalities, we tie our interface to the application. Hence, we have a two input interface that can accomplish tasks more naturally and with greater ease than a single modality.

Conclusions and Future Work

In this work we propose an approach to integrating speech and hand free gestures as two of the most important communication modalities toward a specific task. Integration of multiple modalities into HCII is a natural task, and it is motivated by concurrent use of different modalities in everyday human–to–human interaction. However, until recently, such integration of modalities has been seldom explored. By fusing multimodal features and using the features of each modality to its greatest extent, we hope to provide natural interfaces. Further evaluation of the interface will be done in the forthcoming human subject studies.

References

Baker, P. Battleview. http://www.ncsa.uiuc.edu/Vis/Projects/BattleView.

Cohen, P. R.; Johnston, M.; McGee, D.; Oviatt, S.; and Pittman, J. 1997. QuickSet: Multimodal interaction for simulation set-up and control. In *Proc. of* *the 5th Applied Natural Language Processing Meeting.* Washington, DC: Association of Computational Linguistics.

Hauptmann, A. G., and McAvinney, P. 1993. Gesture with speech for graphics manipulation. *International Journal of Man-Machine Studies* 38(2):231–249.

Kjeldsen, R., and Kender, J. 1996. Toward the use of gesture in traditional user interfaces. In *Proc. International Conference on Automatic Face and Gesture Recognition*, 151–157.

Pavlović, V. I.; Berry, G. A.; and Huang, T. S. 1997. Fusion of audio and visual information for use in human-computer interaction. In *Proc. Workshop on Perceptual User Interfaces.*

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Sharma, R.; Huang, T.; Pavlovic, V.; Schulten, K.; Dalke, A.; Phillips, J.; Zeller, M.; Humphrey, W.; Zhao, Y.; Lo, Z.; and Chu, S. 1996. Speech/gesture interface to a visual computing environment for molecular biologists. In *Proc. International Conference on Pattern Recognition.*