Toward an Intelligent Multimodal Interface for Natural Interaction

Ying Yin
MIT CSAIL
32 Vassar St, Cambridge MA, 02139 USA
yingyin@csail.mit.edu

Randall Davis
MIT CSAIL
32 Vassar St, Cambridge MA, 02139 USA
davis@csail.mit.edu

ABSTRACT

Using a new hand tracking technology capable of tracking 3D hand postures in real-time, we propose to develop a new multimodal tabletop interface for natural interaction. We focus on natural gestures, i.e., those encountered in spontaneous interaction, rather than a set of artificial gestures designed for the convenience of recognition. One camera is used to track the hand. The results of gesture and speech recognition will be fused for mutual disambiguation and co-occurrence analysis. We report on work in progress and the direction of the development.

Author Keywords

Multimodal interaction, natural human computer interaction, gesture recognition, interactive maps.

ACM Classification Keywords

H.5.2 Information interfaces and presentation: User Interfaces

INTRODUCTION

The predominant mode of human computer interaction has not changed substantially since the creation of the windows, icons, menus, and pointer more than thirty years ago. In many ways, they are still convenient and efficient ways for interaction. But with new displays and a growing demand for richer interaction, they appear increasingly limited.

Many new interaction techniques have emerged to provide more natural and convenient modes of interaction, with a common aspiration of making interacting with computation as natural as face-to-face interaction with people. Speech and gestures are the most common modalities for human-human interactions. Hence using them as input provides a natural way for interaction. Naturalness can provide better learnability, flexibility, memorability, convenience and efficiency. With advancement in display technologies, speech recognition and hand tracking, it is increasingly possible to develop intelligent multimodal interfaces that understand what the user is saying and doing, and the user can simply behave.

RELATED WORK

Bolt's pioneering work in the "Put That There" system [2] demonstrated the potential for voice and gestural interaction. In that system, the hand position and orientation was tracked by the Polhemus tracker, i.e., the hand was essentially transformed to a point on the screening.

Multi-touch displays have gained significant media attention and popularity with the introduction of iPhone and Microsoft Surface. Their wide popularity shows the great potential and demand for natural and convenient input techniques. However, with touch-based input, the interaction is still limited in 2D space, and some 3D interaction cannot be realized, or hard and unnatural to specify. For instance, it will be hard to rotate a map in 3D with touch-based display.

Rauschert et al. [6] developed a system called Dialogue-Assisted Visual Environment for Geoinformation (DAVE_G) that uses free hand gestures and speech as input. They recognized that gestures are more useful for expressing spatial relations and locations. Gestures in DAVE_G included pointing, indicating an area and outlining contours. That work, however, mainly tracked hand location, rather than providing both location and posture, as in our work.

INTELLIGENT MULTIMODAL INTERFACE FOR USAR

Emergency response during urban search and rescue (USAR) requires strategic assessment of a large volume of complex information. Intelligent, natural, and multimodal interfaces have the potential to offload the cognitive load from the users, and hence allow them to concentrate more on the decision-making task [7]. Most USAR tasks rely upon geospatial information often presented as maps. As a result, gesture interaction is well matched to this domain. The focus of our work is on developing a multimodal interface for map interaction using the USAR application as a testbed.

System Setup

The custom tabletop structure includes four 1280 × 1024 pixel projectors (Dell 5100MP) that provide a 2560 × 2048 pixel display, projected onto a flat white surface digitizer that is tilted 10 degrees down in front, and is placed at 104cm above the floor [1]. The projected displays were mechanically aligned to produce a seamless large display area. One

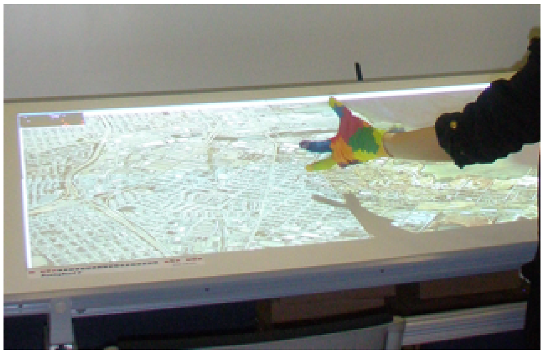

Figure 1. Early prototype of the system.

Fire-i^{TM}digital camera from Unibrain is placed above the center of the tabletop at the same level of the projectors. We use the Google Earth web browser plug-in as our basis for 3D maps.

Hand Tracking

The most common acquisition methods for hand gestures are magnetic trackers, cyber-gloves and vision-based approaches. Acquisition using magnetic trackers and/or cybergloves is efficient and accurate, but suffers from the need to wear restrictive devices. Most existing vision-based hand tracking track only hand movement and finger tips, rather than 3D hand posture [3][4][6]. This limitation often requires that artificial gestures be defined for easy tracking and recognition.

We use the hand-tracking system developed by Robert Y. Wang [8]. With one web camera and one ordinary cloth glove imprinted with a custom pattern (see Figure 1), the system can track 3D hand posture in real-time. It provides rich hand model data with 26 degree of freedom (DOF): six DOFs for the global transformation and four DOFs per finger. The glove is very light-weight, with no additional electronics and wires. As a result, the hand gestures we can use for interaction will not be limited by the hand-tracking hardware, opening up possibilities for developing and investigating more natural gestural interaction.

Gesture Recognition

We adopt the taxonomy of hand movements proposed by Pavlović et al. [5], which distinguishes gestures from unintentional hand movements (like beats) that do not convey useful information. They then further divided the gestures into manipulative and communicative classes.

We are currently working on recognizing manipulative gestures acting on the map. The output from the hand tracker is a time sequence describing the joint angles and the 3D position and orientation of the hand. We use Hidden Markov Models (HMMs) to train and classify gestures. For direct map manipulation, we have trained gestures including: panning, rotation about x, y, and z axis, and zooming. We are refining the algorithm to improve the accuracy of classifying continuous gestures.

Fusion of Speech and Gesture

We will use an off-the-shelf speech recognition engine for keyword spotting, then combine speech and deictic gestures. In one example scenario, during an earthquake relief effort, the commander at the command center asks the system to show all the hospitals in a certain area by asking aloud “What hospitals are there in this area?” while using her hand to circle out the area. She then asks the system to show the specific information about one of the hospitals by pointing at it, and then indicates a route for directing casualties to the hospital.

The challenge for this part is the speech and gesture alignment and the fusion of two modalities. We will use separate modules for gesture and speech recognition, and a probabilistic co-occurrence analysis module to combine speech and gesture.

FUTURE WORK

After finishing the prototype, a user study will be conducted to evaluate the effectiveness of multimodal interaction versus traditional mouse and keyboard interaction. We also plan to incorporate stereo imaging or a time-of-flight camera to get more accurate data on the height of the hand relative to the table so that the table can become touch sensitive too. In this way, we can create an interaction environment with a rich variety of input modalities: high resolution input from the stylus on the digitizer table, low resolution input from the finger tips, 3D hand gestures and speech.

REFERENCES

1. Ashdown, M. and Robinson, P. Escritoire: A personal projected display. *IEEE Multimedia 12, 1* (2005), 34–42.
2. Bolt, R. A. “Put-That-There”: Voice and gesture at the graphics interface. In *SIGGRAPH ’80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, ACM (1980), 262–270.
3. Demirdjian, D., Ko, T., and Darrell, T. Untethered gesture acquisition and recognition for virtual world manipulation, 2003.
4. Oka, K., Sato, Y., and Koike, H. Real-time fingertip tracking and gesture recognition. *IEEE Computer Graphics and Applications 22, 6* (2002), 64–71.
5. Pavlovic, V. I., Sharma, R., and Huang, T. S. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19* (1997), 677–695.
6. Rauschert, I., Agrawal, P., Pyush, I. R., and Sharma, R. Designing a human-centered, multimodal GIS interface to support emergency management, 2002.
7. Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., Maceachren, A., and Sengupta, K. Speech-gesture driven multimodal interfaces for crisis management, 2003.
8. Wang, R. Y. and Popović, J. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics 28, 3* (2009).