Toward an intelligent multimodal interface for natural interaction

Ying Yin **Randall Davis**
Computer Science and AI Lab, MIT
{yingyin, davis}@csail.mit.edu

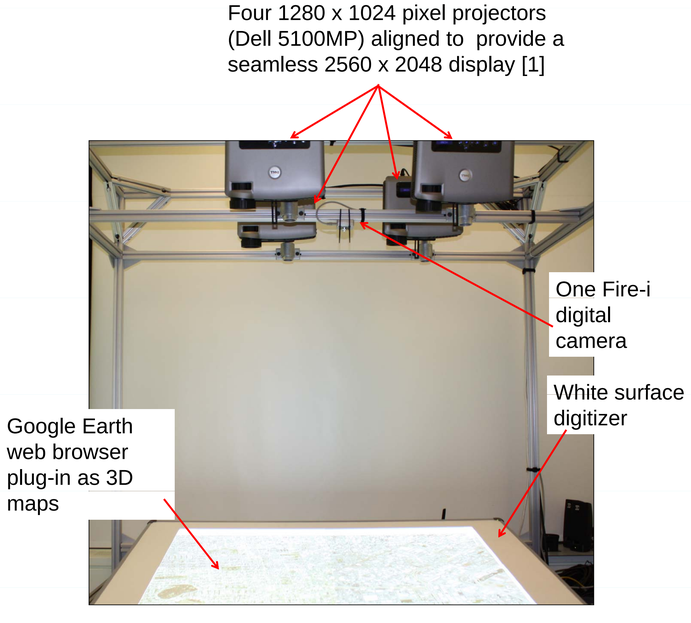
Introduction

We are developing a new multimodal tabletop interface for natural interaction, based on a new technology capable of real-time tracking of 3D hand postures. 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. We will use an off-the-shelf speech recognition engine for keyword spotting, then combine speech and deictic gestures.

We have chosen urban search and rescue (USAR) as our application domain, because most USAR tasks rely upon geospatial information, typically presented as maps. Gesture interaction is thus well matched to this domain.

Figure 1. Tracking the hand in 3D enables more flexible and natural gestures

System Setup



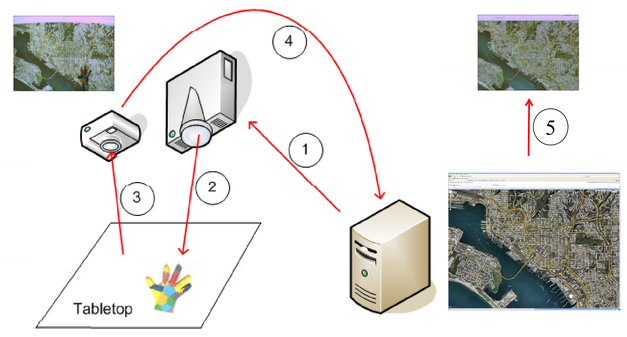
Hand Tracking

We use the hand-tracking system developed by Wang [2], which is capable of tracking 3D hand posture in real-time, providing hand model data with 26 degrees of freedom. The system uses a simple web camera; the user wears a cloth glove imprinted with a custom pattern. The glove is light-weight, with no additional electronics and wires. Hence, the gestures we can use will not be limited by the hardware.

The hand-tracking software was developed originally for use in an office environment with standard illumination, where the gloved hand stands out from the simple and relatively static background. Our context, however, requires working with maps that are both complex and dynamic. We want to remove the effects produced by projecting the maps: the complex background and the non-uniform, dynamic illumination of the hand.

To do this, we dynamically compute the background image as seen by the camera, transforming the map image step by step to simulate the effects of the projector and the camera.

Schematic showing the transformation of image pixels



1. Map image sent from computer to projector
2. Non-linear color change when light emitted from projector
3. Non-linear color change when light received by camera, and scaling, cropping, and distortion by camera
4. Image including the hand sent from camera to computer
5. Apply geometric and color transformations to map image to compute background

We divide the camera image (pixel by pixel) by the background image computed above, removing the background and correcting the color of the glove.

Denoised camera image → Background eliminated image → 3D posture model [2]

Gesture Recognition

We divide gestures into two classes: manipulative and communicative.

We are currently working on recognizing manipulative gestures acting on the map. For direct map manipulation, we have trained gestures including pan, pitch, roll, yaw, and zoom.

Examples of gestures

Pitch Roll Zoom

The output from the hand tracker is a time sequence (12fps) describing the joint angles and the 3D position and orientation of the hand. We use Hidden Markov Models (HMMs) to train and classify gestures. To date we have achieved 95% accuracy on isolated gesture recognition, and are working on continuous gesture recognition.

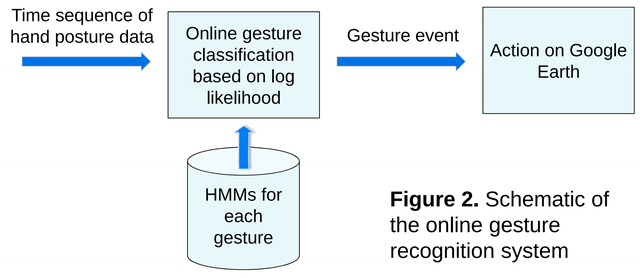


Figure 2. Schematic of the online gesture recognition system

Future Work

We 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, enabling a variety of touch sensitivity. This will permit a rich interaction environment with a variety of input modalities: high resolution input from the stylus on the digitizer table, low resolution input from the finger tips, 3D hand gestures, touch sensitivity, and speech.

References

1. Ashdown, M. and Robinson, P. Escritoire: A personal projected display. IEEE Multimedia 12, 1 (2005), 34–42.
2. Wang, R. Y. and Popović, J. Real-time hand-tracking with a color glove. ACM Transactions on Graphics 28, 3(2009).