MULTIMODAL INTERACTIVE SPACES: MAGICTV AND MAGICMAP

MARCELO WORSLEY AND MICHAEL JOHNSTON

STANFORD UNIVERSITY, AT&T LABS RESEARCH

ABSTRACT

Through the growing popularity of voice-enabled search, multimodal applications are finally starting to get into the hands of consumers. However, these applications are principally for mobile platforms and generally involve highly-moded interaction where the user has to click or hold a button in order to speak. Significant technical challenges remain in bringing multimodal interaction to other environments such as smart living rooms and classrooms, where users speech and gesture is directed toward large displays or interactive kiosks and the microphone and other sensors are 'always on'. In this demonstration, we present a framework combining low cost hardware and open source software that lowers the barrier of entry for exploration of multimodal interaction in smart environments. Specifically, we will demonstrate the combination of infrared tracking, face detection, and open microphone speech recognition for media search (magicTV) and map navigation (magicMap).

Index Terms— multimodal integration, open microphone, speech recognition, gesture recognition

1. INTRODUCTION

Interfaces supporting rich multimodality where users can combine speech and gesture are making their way from research prototypes to consumer applications. For example, Speak4it[SM] [1] is a mobile application allows users to simultaneously speak local search queries e.g. "gas stations" while drawing on a map to indicate their region of interest. In addition to their use on mobile devices, multimodal interfaces also have significant potential applications in smart environments where users interact with large screen displays. Examples include searching for and interacting with in-home entertainment systems and interaction with a digital blackboard by students in a classroom. In these settings, user gestures involve hand movements towards a distant screen display and, ideally the system will always be listening and users will not have to be instrumented with click to speak buttons in order to indicate when they are addressing the system. There has been a significant body of previous work on multimodal interaction for smart environments, e.g. [4], though a lot of work has focused more of meeting scenarios rather than interaction in a living room setting [2]. Also, much of this work as involved a significant amount of costly hardware, multiple cameras, microphone arrays, other sensors, and multiple servers to capture all of the sensor input. In this work, we explore the creation of an environment for exploration of multimodal interaction that utilizes low cost commodity hardware combined with a single Linux server utilizing open source software. Specifically the modalities supported are computer vision, for face detection, infrared tracking, for gesture input, and speech input. The environment operates in an 'open sensors' mode where all of the sensors are continuously receiving and processing data. In the following sections, we describe the hardware, the software framework, and then go to describe to demonstration prototypes magicTV and magicMAP built using the framework.

2. HARDWARE – LOW-COST SENSORS

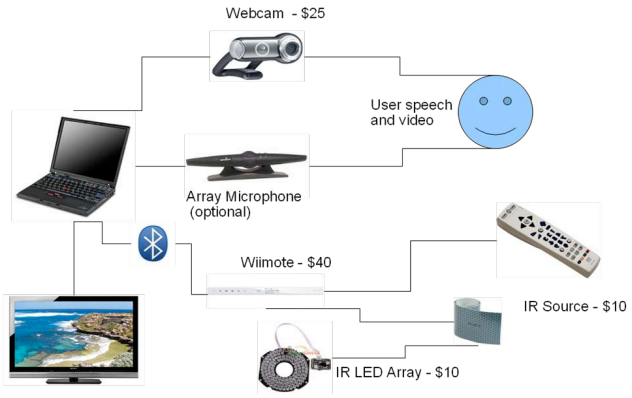

Figure 1 - Hardware Architecture

In addition to a computer and monitor, the main hardware components are for capture of video, speech, and gesture. For video, we employed a low-cost high definition webcam. This captures the scene in front of the system and is used for person and face detection. For speech capture, we utilize an array microphone. In order to provide precise pointing, and avoid problems with lighting conditions, rather than using computer vision for gesture capture we chose to use infrared tracking. Specifically we leverage the 1024x768 infrared camera in the Nintendo Wiimote. Note here that the user does not hold the Wiimote, rather it is mounted below or above the screen display and is used as a sensor. The IR camera is capable of capturing the location of up to 4 simultaneous IR sources within its field of view. The demonstration supports the use of a variety of IR sources. We experimented with three different kinds of IR source: 1. Infrared LED Pen held by the user, 2. IR array that circles the Wiimote, 3. Traditional TV remote. For 2., we place high gain reflective tape on the users hands and the sensor

tracks light reflected from the IR array [3]. While the framework supports all three kinds of IR inputs, in testing we found the traditional remote to be the easiest and most natural to use. Users are familiar with holding the remote, and this framework allows them to use any IR remote as a pointing device ('magic wand') in order to make gestures toward the display. Because of the distinct properties of each IR source, the application framework supports different settings to ensure usability and low latency regardless of the IR source's limitations.

3. SOFTWARE – A CROSS-PLATFORM MULTIMODAL APPLICATION FRAMEWORK

Our goals for the software framework were that it should be cross-platform, support capture and coordination of the three input modalities within a single language, and enable specific applications to be developed using HTML and Javascript. The system is built in Python and brings together a variety of Python libraries to support processing of the multiple data streams: *pyaudio* – for capturing streaming audio; *opencv* – for capturing and processing streaming video data; and *pywii* – a SWIG wrapper on the wiiuse library that communicates with the Nintendo Wiimote (Figure 2). By leveraging the above libraries and a series of custom algorithms, the software is able to do face and people recognition, basic shape recognition (lines, polygons and circles), and dwell detection.

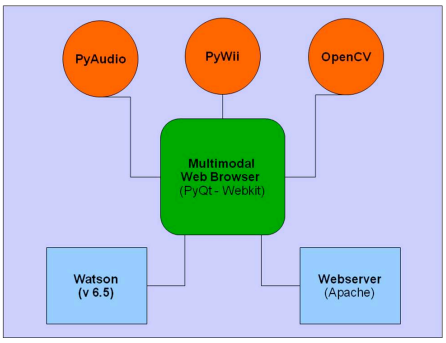

Figure 2 – Software Architecture

In addition to these three libraries, the application framework is built on PyQt4. PyQt4 contains a variety of classes that permit rendering of web content, direct manipulation of DOM objects, application threading and controlling interaction with the graphical user interface. Applications, such as the demonstrations described in the next two sections, are built using HTML and Javascript and are rendered using the webkit browser within PyQt. The application framework also incorporates tight coupling with the AT&T Watson speech recognizer system [4] and a local Apache webserver for serving the HTML content.

4. MAGICTV – IN-HOME MEDIA SEARCH

The first demonstration shows how speech, gesture, and face detection can be used together in a system for in-home media search. This application extends the VideoMatch system [5]. Users can use speech to search for media, e.g. "new episodes of reality TV shows". Rather than paging through results using up and down buttons, the user can simply point at results and other screen elements using the remote in order to get more information. Gesture recognition is used for pagination, the user simply makes a swipe gesture up or down in order to scroll the results. The system operates in an open microphone mode. In order, to detect whether speech is addressed to the system, the application monitors the results of face detection and recent gesture activity in order to determine if speech is addressed to the system. Figure 3 shows a user using the magicTV system and highlights the various hardware components.

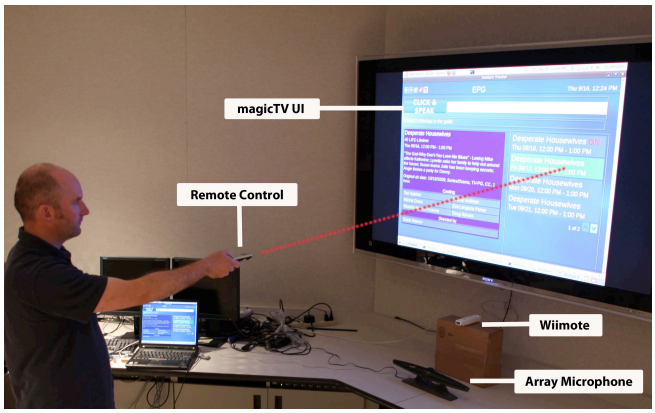

Figure 3 – magicTV prototype

5. MAGICMAP –INTERACTIVE NAVIGATION

The magicMAP prototype brings multimodal map-based interaction into the living room - possible uses include planning family vacations or business trips. In addition to voice commands for navigation e.g. "show san Francisco California". The system supports direct manipulation of the display using gesture input. The user can 'grab' hold of the map and pan it around using gestures made in space using the remote control. The system also supports multimodal integration, for example users can circle a region of the screen using the remote and say "zoom in here".

REFERENCES

[1] http://speak4it.com/

[2] http://www.amiproject.org/

[3] http://johnnylee.net/projects/wii/

[4] Corradini, A., R. Wesson, and P. Cohen. 2002. A Map-based System using Speech and 3D Gestures for Pervasive Computing. In Proceedings of ICMI 2002. pp. 191-196.

[5] Goffin, V., C. Allauzen, E. Bocchieri, D. Hakkani-Tur, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar. 2005. The AT&T WATSON speech recognizer. In Proceedings of ICASSP. pp. 1033–1036.

[6] Johnston, M., L-F. D'Haro, M. Levine, B. Renger. 2007. A Multimodal Interface for Access to Content in the Home. Proceedings of ACL 2007. pp. 376-383.