

Multimodal Interactive Advertising

D.M. Lyons, D.L. Pelletier, D.C. Knapp

Adaptive Systems Department

Philips Research Briarcliff Manor NY 10510

{dml,dlp,dck}@philabs.research.philips.com

Abstract

An implementation of a multimodal interactive advertising system is described in this paper. The system integrates computer vision and speech recognition technologies to allow the advertising display to determine when a customer is nearby and to respond to that customer's location in space, gestures and speech. The advertising application is described here in detail. The system architecture and hardware/software modules involved are presented in overview. The application has been tested in the laboratory and exhibition hall settings but has not yet been deployed in public.

1. Introduction

Two of the most common user interface devices today are the remote control and the keyboard. From the perspective of the consumer, neither of these is completely natural to use -- they take preparation and some training -- and they can rarely be called a pleasure to operate. This can prevent a consumer from appreciating the full functionality of a device. We are all familiar with the story of how few people actually set their VCR clock. Indeed, this may even sour a consumer completely on a brand name, since the user interface is the main channel through which a consumer experiences a product. We take the approach that to make a device natural and a pleasure to use, it needs to use similar input channels to those that people use -- their sight, their hearing and so forth. The consumer can "interact" with a device simply by walking up to it, and/or by pointing and/or by speaking with it and so forth.

This paper describes an implementation of an advanced user interface for a public advertising system for a department store or shopping mall. The interface was designed to allow a passing customer to interact with the system in an almost unconscious fashion. Computer vision and speech recognition are integrated into a multimodal user interface that allows the system to sense when a customer is nearby and to respond to that customer based on his/her body position, gestures and speech.

The paper is laid out as follows. Section 2 contains an introduction to the literature in this area. Section 3 describes the Advertising application in detail. Section 4 overviews the system architecture and the modules involved. Section 5 concludes with our evaluation of the performance of the system and pointers for further work.

2. Literature Survey

The MIT ALIVE system [1] demonstrated the use of computer vision as an integral part of the user interface. In the ALIVE application, a person entered the "interaction space," an open area in front of a large projection display screen, and could then interact by gesture and motion with graphical animated creatures. A computer vision system, [2] connected to a single camera placed over the interaction space, extracted the gesture and motion information. More recently, Lucente [3] has applied this approach, integrated with speech recognition, to a scientific visualization application in his "Visualization Space."

A visualization application using computer vision is also described by Pavlovic et al. [4]. However, in that work, the interaction area is limited to a small volume in which a person places his/her hand. Two cameras are used to get simultaneous top and side views of the hand. Finger gestures can then be used to control the motion of 3D objects shown in the MDScope virtual reality environment.

Moving further away from the concept of a defined "interaction space," Kahn and Swain [5] describe the Perseus system, a gesture based interface for a mobile robot. Perseus interprets pointing gestures to direct a mobile robot to proceed to the indicated location. Christian and Avery [6] describe the Digital Smart Kiosk work. In that application, computer vision is used to direct the gaze of an animated human head so as to follow a person walking by, or interacting with, the kiosk. Turk [7] has described a similar application, but for a desktop scenario, in which computer vision provides information to enrich the "Peedy" character's interaction with a user.

Our application, interactive advertising, is most similar to the Digital Smart Kiosk application in that it involves use by the general public. However, like the Visualization Space work, we also need to interpret communication gestures and coordinate with speech recognition.

3. Multimodal Interactive Advertising

The *Interactive Advertising* application is targeted at allowing a customer in a department store or mall to interact with an advertising sequence running continuously on a large screen TV. The apparatus used (Figure 1)

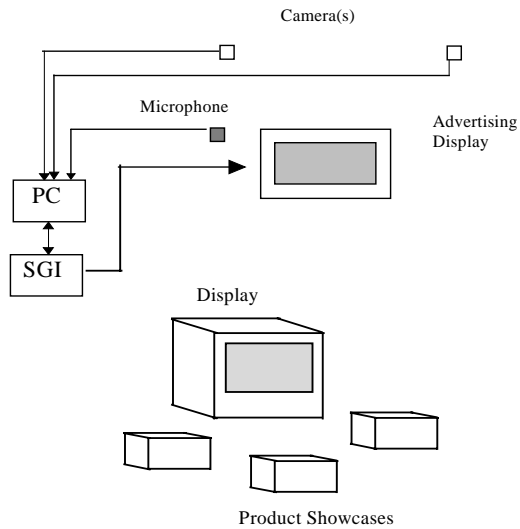


Figure 1: The Equipment used for Multimodal Interactive Advertising



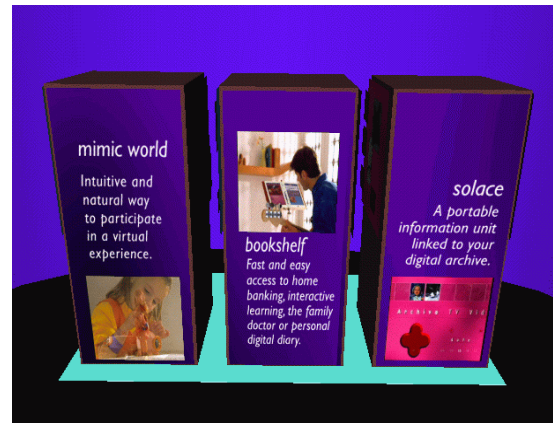
Figure 2: Advertising Display with no-one in view.

consists of a large screen display with either one or two cameras mounted on the top of the display. (Once positioned, the cameras can be automatically calibrated so that accurate visual tracking can be conducted.) A microphone is positioned close to the display. The cameras and the microphone are both connected to a PC, which does the visual and speech analysis. An SGI O2 is used to control the advertising content. The PC and SGI communicate over a serial link.

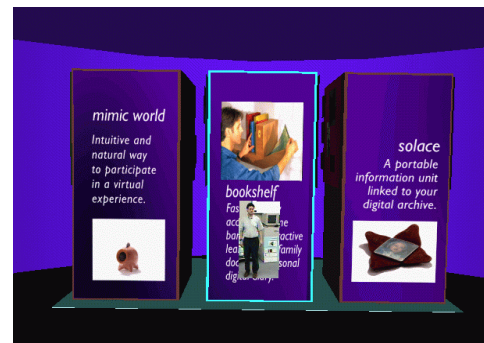
The equipment is set up in the department store, movie theater lobby, etc., as follows: the products being advertised (three products in our example) are placed in showcases just in front of the TV. When no one is in view, the TV runs displays on all three products (Figure 2; the “pictures” on each cube are actually movies playing on the cube).

3.1 Approaching the Advertising Display

When a customer walks into the view of the cameras, the system adapts the advertising to that of the product stand nearest the customer. As an *attention-getter* the tracked video of the person is pasted onto



the advertising. In this example, that means a section of the video corresponding to an extended bounding box is overlaid on the advertisement for the product nearest which the customer is standing (or passing; figure 3). A



more sophisticated example would be that for a hat advertisement, where a sample hat could be pasted onto the head of the person.

3.2 Looking at a Product

Once the display has gained the customer’s attention, a likely next natural action for the customer is to walk over and look at one of the product showcases. When the customer walks closer to a product stand, the advertising is again altered to show a full screen movie of the product on *that* product stand (Figure 4). The point is that the person is controlling the system by doing nothing other than walking around naturally. A more conventional user interface would have a menu on the screen: *To see the movie for product 1, press 1 and enter; for product 2, press 2 and enter*, etc. The vision-based approach eliminates this barrier between the user and the machine; the machine can directly determine in what product the user is interested.

3.3 Getting Product Information

A customer can more actively control the information. If

you watch the product movie for a while, a “flash card”



Figure 4: A full-screen product movie is displayed

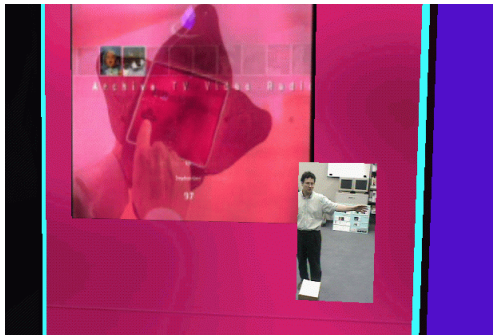


Figure 5: Getting additional information using gestures

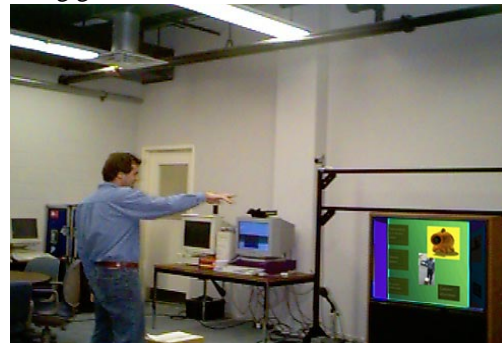
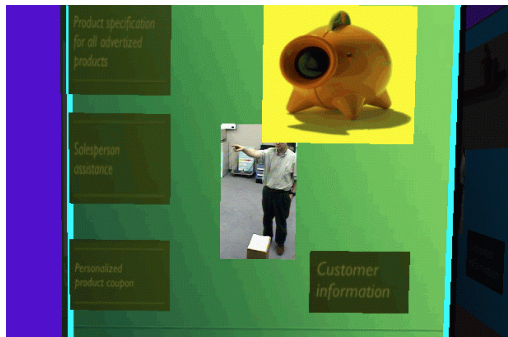


Figure 6: Selecting from a menu by pointing and speaking, screen & side view.

pops up to tell you that you can access more information by stretching out your right arm. Stretching out one’s hand to attract the system’s attention is arguably a reasonably natural gesture. The cube on which the information is presented rotates clockwise (Figure 5) when the customer raises his/her right arm. (It will rotate counter-clockwise for the left arm). Note that the tracked video image of the person is composited onto every page, so that the customer is always aware of who is in control.

3.4 Making selections with Pointing & Speaking

If the customer again raises his/her right hand, indicating the desire to have more information about the product, the cube again rotates counter-clockwise and presents a new display (Figure 6) which has a graphical menu. Again, a yellow “flash card” appears on the screen informing the customer that pointing at an item will highlight that item;

and speaking the word “Okay” will select the highlighted item. A customer can make selections from items on the screen by pointing at them and using speech recognition to confirm his/her choice. Pointing is simply that: either hand can be used and no real precision on the part of the customer is necessary as the system of highlighting the selection provides feedback (Figure 6).

4. System Architecture

The Multimodal Interactive Advertising implementation has two main modules:

1. The *MIA Display Module*, running on the O2, sequences and renders the advertising presentation based on user input supplied by the perception module.
2. The *MIA Perception Module*, running on a 200MHz PC, analyzes the camera input to determine body position and gesture information and runs a speech recognition program.

Communication between the two modules is via a 9600 baud serial link. The MIA display module is programmed in SGI Inventor and uses the SGI DMBuffer library to play movies and video on 3D objects. Currently, the display module also runs the Gesture Interpretation module, *Gesture Pipes* [8].

The Perception Module runs both the Speech Recognition and the Computer Vision modules. Integration of speech and vision is done in the *Gesture Pipes* module.

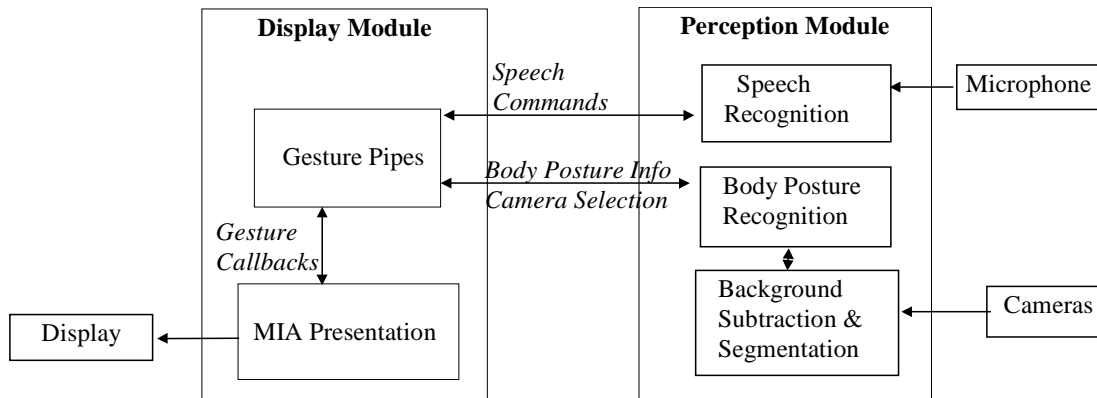


Figure 8: MIA System Architecture

The computer vision modules consist of

1. a higher-level body posture module that labels body features such as hands and legs and determines their position in 3D space, and
2. a lower-level segmentation and background subtraction module.

Background subtraction is done by sampling the view when no customers are around, and subtracting this information from every subsequent image. The lower-level module uses the functionality of the Philips Single Board Image Processor, a commercial frame-grabber with some DSP functionality.

5. Conclusion

The Multimodal Interactive Advertising system has been tested under laboratory conditions and five days of exhibition usage. It has not been tested by exposure to the general public (as for example, the Digital Smart Kiosk has [6]). Overall reaction to the interface was very positive, with many remarking on the “magic” nature of the control. However, several flaws were also detected:

1. Handling crowds. The current system uses static background subtraction and selection of the largest silhouette as the target user. This only works well under uncrowded conditions. It was necessary to strictly control the flow of people by the system, effectively to one person at a time, to prevent confusing the system.
2. Coarse gestures. While pointing was universally appreciated as a useful gesture, using side arm gestures was faulted as being too awkward in public. Smaller control gestures would be preferred.
3. Gesture interpretation issues. Few false positives were experienced with the gesture interpretation code, however, the speed of the pointing interpretation was seen as a flaw. The response of change of selection to hand motion varied between one and five seconds.

In future work we will address these problems. We plan to address the speed issue with a redistribution of functionality between the two machines: the PC is currently

underutilized, while the O2 is overloaded. The gesture pipes software will be transferred to the PC. This has the other beneficial effect of reducing the amount of information that needs to be transmitted on the serial link between the machines.

Handling crowds better will require us to make more extensive use of human body models, something the system does very minimally now.

Allowing for fine control gestures may be more problematic, as it may require the use of a PTZ camera to image hand motions at the level of detail used by [4].

References

1. Maes, P., *Artificial Life meets Entertainment: Lifelike Autonomous Agents*. CACM, 1995. **38**(11): p. 108-114.
2. Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A., *Pfinder: Real-Time Tracking of the Human Body*, Media Lab Report 353. 1996, MIT Media Lab: Cambridge MA.
3. Lucente, M. *Visualization Space: A Testbed for Deviceless Multimodal User Interface*. in *AAAI Spring Symposium Series*. 1998. Stanford CA.
4. Pavlovic, V.I., Sharma, R., and Huang, T.S. *Gestural Interface to a Visual Computing Environment for Molecular Biologists*. in *Face and Gesture Recognition*. 1996.
5. Kahn, R.E., and Swain, M.J. *Understanding People Pointing*. in *Int. Symp. on Computer Vision*. 1995.
6. Christian, A.D., and Avery, B.L. *Digital Smart Kiosk Project*. in *CHI'98*. 1998. Los Angeles.
7. Turk, M. *Visual Interaction with Lifelike Characters*. in *Face & Gesture Recognition*. 1996.
8. Lyons, D.M., and Murphy, T.G., *Gesture Pipes: An Efficient Architecture for Vision-Based Gesture Interpretation*, PRB-TN-97-039. 1997, Philips Research: Briarcliff NY.