ABSTRACT

Human Computer Interface (HCI) in applications for the maintenance of complex machinery such as an aircraft can be enhanced by exploiting new developments in HCI. We have developed a multimodal HCI demonstration system for maintenance applications, incorporating Augmented Reality (AR), Speech Recognition, and 3-dimensional audio technologies. The Augmented Reality interface is based on an original dynamic tracking approach to provide rapid update of the scene with graphical overlays. We enhance the use of this interface with speech recognition to control the system and to add annotations using dictation-based text information. A combination of 3-D audio, graphic animations, and text displays is used to communicate information to the user.

INTRODUCTION

Maintenance tasks typically require the maintenance worker to repeatedly refer to documentation about the machinery being serviced. This documentation typically contains graphical elements such as a layout diagram of the machine, and textual elements such as a description of the procedure for replacing a machine part. The act of referring back and forth between the machine and the documentation could be tiring, or even difficult to execute in cramped spaces. One of the solutions that has been proposed for addressing this scenario is to show the documentation in a display that is suspended in front of one or both eyes (Figure 1). These are typically head-mounted displays (HMD) and are either see-through or opaque.

The advantage of HMDs is that information can be presented to the user while the actual scene is being viewed, without taking the eyes off the scene. This in itself will be a useful display configuration. In fact, such a display can be enhanced using technology developments in the field of Augmented Reality (AR), in combination with input and output modalities such as Automatic Speech Recognition (ASR) and 3-D audio presentation. In this paper, we describe a system that incorporates these components in a demonstration prototype. In this demonstration prototype, the user can view CAD models of an object displayed “in place” on top of the object, as if the CAD model is attached to the real object being viewed. This capability is possible due to the use of a new “dynamic registration” technique that is outlined in this paper. The user can query about parts of the object using speech commands (“Where is the power supply?”), and the response is presented as a graphical animation, textual display, or a 3-D audio icon as appropriate.

Figure 1. Head-Mounted display for maintenance tasks.

The integration of the various components (AR, ASR, 3-D audio) is carried out in a client-server architecture. This architecture permits rapid reconfiguration of our prototype system to exploit new advances in ASR and 3-D Audio technologies.

A related technology, under the theme “outdoor augmented reality,” is aimed at displaying registered terrain information in the flight deck. This will enable the pilot to see on a Heads Up Display (HUD) the actual location of natural and man-made structures and
weather events. This technology is also being investigated at Rockwell Science Center, and has been published [2]. In this paper, we start with a presentation of the user interaction capabilities of the system, then describe the various components, and conclude with a discussion.

PROTOTYPE SYSTEM DESCRIPTION

AUGMENTED REALITY INTERFACE

The Augmented Reality (AR) interface allows a viewer to view a CAD model of an object superimposed on top of the real object. Figure 2 shows an example of this type of display in which a CAD model of a PC (tower chassis) is superimposed on live video frame from a camera that observes the PC. The resulting display provides virtual “X-ray” vision to the user. In this display, the CAD model and other information have been superimposed directly on the incoming video stream. This is one way of presenting the information. An alternative way will be to display only the CAD model in a display that is placed in front of the user’s eye(s), as shown in Figure 1. In this case, the superposition will be carried out perceptually.

Figure 2. Augmented Reality display with a CAD model of a PC superimposed on live video frame containing the PC. Object components may be highlighted as shown.

The display is dynamic. In other words, the user could walk around the object, look up or down, or perform other maneuvers that alter the viewing position and orientation relative to the object. The system will continue to track the user’s viewpoint relative to the object and will keep the display updated – i.e, the superposition will be in the right place. Thus, the information displayed will always be not only just in time but also just in place. Additionally, if the user looks away from the object and looks back at it again, the system will automatically synchronize the CAD model with the object in view.

If the user has to deal with multiple objects whose locations are known, by tracking the location of the user, the system could automatically determine which of the objects are in view and display their CAD models in registration with the view. This is not implemented in our prototype but can be readily added by using a suitable tracking mechanism.

SPEECH INTERFACE

The speech interface can be used to perform any of the following:

- Query the location of a specific component
  - Sample: “Where is the CDROM drive”
  - Response: CDROM drive is highlighted by a blinking animation. If the component is outside the field of view a 3-D audio icon is played (see next subsection)
- Display control information
  - Sample: “Show the frame rate”
  - Response: Frame rate displayed and played out using Text-to-Speech (see below)
- Manipulate annotations
  - Sample: “Add a virtual note”
  - Response: An annotation icon is displayed in 3-D attached to the currently selected object
- Dictate annotations
  - Sample: “This component has to be replaced”
  - Response: The dictated text is displayed and the annotation icon will be visible, attached to the currently selected object

The first three types of speech input in the above list are speaker-independent, and are usually very accurate. For obtaining acceptable dictation recognition results, it is preferable that the speech recognition system be trained by the user.

In addition to command and dictation recognition, text-to-speech (TTS) is used to provide audio alerts and responses to the user.

3-D AUDIO INTERFACE

When augmented reality information to be displayed is outside the field of view, we can appeal to a modality that is not restricted by the visual field of view, namely audio. Audio icons are not new; they are customarily used in applications to alert the user on error events. A new audio capability that is relevant to a user dealing with spatial information is 3-D audio. With dramatic reduction in 3-D audio hardware costs, it is now possible
to provide information in the form of 3-D audio inexpensively.

In our system, when the user queries about an object component that is outside the current field of view, (such as “where is the printer?” and the printer is not within view), a 3-D audio icon is played to direct the attention of the user to the component. We have experimented with two different ways of presenting the information:

- Playing the audio icon at the static location that corresponds to the component.

- Playing a moving audio icon that starts from a canonical position in front of the user and moves to the component location.

The latter appears to be more compelling than the former, but more rigorous experiments must be carried out to verify this empirical result. More research is also needed to determine the most effective motion paths for the audio icon.

**PROTOTYPE SYSTEM COMPONENTS**

The overall schematic of the system is shown in Figure 3. The three main system modules, namely the AR, ASR, and 3-D audio are implemented as separate processes that reside on a network and communicate through TCP/IP sockets. Speech recognition and 3-D audio are encapsulated in servers with a serial-like ASCII protocol. This architecture allows us to continuously improve the different modules when better hardware (new 3-D audio boards) and software (better speech recognition engine) solutions become available. In our research environment, this architecture also allows us to exploit speech recognition and 3-D audio services from multiple platforms (Windows 95 and NT, SGI IRIX, and UNIX). The AR application is the main process, and is the client for the speech and 3-D audio servers.

**AUGMENTED REALITY INTERFACE**

In order to display the CAD model of an object in such a way that the model appears to be superposed on the real object in the field of view, the model has to be registered with the view. This can be accomplished by tracking the viewpoint of the user relative to the object, using images from a camera that is close to the viewpoint of the user. Viewpoint tracking using video-based methods is passive, and can achieve high accuracy in the alignment of graphical information display with the view of the real world. The tracking in our system is based on *Visual Servoing* [1]. Visual servoing is to control a system - typically, a robot end-effector with a camera – based on processing visual information. In this approach, the error between measured image positions of visual features and their predicted image positions is minimized with respect to the camera motion parameters. The camera is moved in closed loop using the calculated motion parameters to reduce the error measure, as shown in the flow diagram in Figure 4. The images from a video camera are processed to measure feature positions, and the “virtual” camera (which renders the graphics) is controlled.

![Figure 3. Schematic of the system showing the AR, ASR, and 3-D audio components.](image-url)

![Figure 4. Flow diagram for the AR module.](image-url)

We use concentric ring fiducial markers as features. These markers were chosen for easy detection in clutter and under a wide range of viewing angles. Each marker
has a unique ring structure for identification. The rings form a binary pattern which corresponds to a numeric ID value. The currently used version with four rings allows ID numbers from 0 to 15 (4 bit).

We believe that our closed-loop, vision-based head tracking method [3] solves the visual alignment problem directly, and is robust due to the use of specially constructed markers (Figure 5).

Figure 5. Fiducial markers used for tracking. These have a radial binary encoding and are symmetric from a wide range of viewing angles.

The system speed and capabilities of the rendering engine set the limits for the complexity of information that can be displayed. In our prototype, we display wireframe and Gouraud-shaded 3-D objects and colored text. The graphical display is overlaid on to the live video stream, using the head tracking technique described in Section 2.1. The result is a form of “X-ray” vision that enables the user to visualize internal components of the viewed object. We use simple animations (flashing) to indicate selection when the user queries about a specific component. Textual annotations (see Section 2.3) can be displayed for the selected component.

Figure 6. AR display with camera positioned at different locations around the object.

AUTOMATIC SPEECH RECOGNITION (ASR) SERVER

Rockwell Science Center’s Automatic Speech Recognition (ASR) Server software provides an easy way to rapidly prototype speech-enabled applications regardless of the computing platform(s) on which they execute. The ASR server provides both automatic speech recognition and text-to-speech synthesis (TTS) capabilities. The ASR capability is obtained through abstraction of a commercially available off-the-shelf speech recognition technology, IBM ViaVoice. The TTS functionality provided with the ViaVoice engine is likewise abstracted and exposed to client applications. The ASR server’s architecture provides for the future addition of other vendors’ speech recognition technologies as needed. A client application connects to the ASR server over an IP network using TCP sockets. Although the ASR server runs on a Windows 95/Intel Architecture PC, the client applications may run on any operating system that supports TCP/IP networking.

In the AR system, we use speech to control the system (e.g., “Start the demo”), to query the state of the system (e.g., “What is the frame rate?”), to query the location of components (e.g., “Where is the power supply?”), and to dictate “virtual notes” that can be attached to the components and retrieved at a later time. The TTS is used to provide informative audio responses and alerts to the user (e.g., “Please repeat command”).

3-D AUDIO SERVER

A three-dimensional (3D) audio system provides an auditory experience in which sounds appear to emanate from locations in 3D space. Our particular interest is in the application of HRTF (Head-Related Transfer Function [*4])-based 3D audio to provide cues about objects outside the field of view. Our 3D audio server is built upon an API for a chip from Aureal that implements Microsoft DirectSound3D functionality and hardware HRTF processing. The server is based on TCP/IP sockets. The sound source signals are stored as wave files (.wav) on the server. A client application can connect to the server, designate the sound source signals to be played, and stream position/orientation information for the listener and objects to which the sounds are attached in 3D. The server operates at 30 frames per second (for position and orientation updates) and can simultaneously play up to three 44.1 kHz spatialized sound sources.

SYSTEM INTEGRATION

All the system components of the previous section have been implemented on PCs running Windows 95 or NT. A schematic of the system is shown in Figure 3. The servers (ASR, 3-D audio) are hosted on a Windows 95 platform with a 300 MHz Intel Pentium II processor, and the AR application (visualization and image processing) is implemented on a Windows NT machine with a 200 MHz Intel Pentium processor. The PC is equipped with an Imaging Technology color frame-grabber that
digitizes video signal from a Cohu 2200 CCD camera. The image processing is carried out on the CPU, to detect and identify fiducial markers (Section 2.1). The AR visualization is implemented as a World Tool Kit (WTK) application. All user interactions are carried out in the main action loop of the WTK application.

The AR technology is demonstrated by superimposing the WTK graphics on top of incoming video signal (see Figures 2 and 5). The display can be easily modified for a see-through system. Using a speech command the user can ask the system to highlight a specific component. If the component is within view, the selected component is flashed on the display. If the component is outside the field of view, a 3-D audio icon is played and moved towards the location of the component (from a canonical location in front of the user) to direct the user's attention. Using dictation recognition, the user can compose a virtual note and attach it to a component. Previously attached virtual notes may be viewed or deleted using speech commands.

DISCUSSION

We have developed a prototype system to demonstrate multimodal interaction in a maintenance application, using Augmented Reality, Speech Recognition, and 3-D audio technologies. With this prototype system, a user can move a camera around an object whose CAD model is available. A video display enriched with aligned 3-D graphical information is presented to the user and is updated continuously as the camera is moved. In the current implementation (on a 200MHz Pentium processor), we obtain 6-10 frames per second. The user can control the system using speech commands; add annotations using dictation; and be presented with 3-D audio icons to direct attention to objects outside the field of view.

This prototype demonstrates the applicability of multiple modalities to a maintenance application. There are several challenges in turning this prototype into a product: computational complexity of video processing, lack of high quality small form factor framegrabbers, lack of inexpensive high resolution head-mounted displays, and need for reliable indoor position trackers (like GPS for outdoors). In addition for this technology to be widely applicable, techniques object recognition and tracking in video sequences need to mature. Recognizing objects from images is in the general case an unsolved problem. Techniques for tracking have been more widely developed, but reliable tracking under varying light conditions, clutter, and occlusions is still a topic of active research.

CONTACT

The address for all authors is: Rockwell Science Center, 1049 Camino Dos Rios, Thousand Oaks, CA 91360.

Electronic Mail addresses are:

Vsundar@rsc.rockwell.com
Rwbehringer@rsc.rockwell.com
Slchen@rsc.rockwell.com
Kkwang@rsc.rockwell.com

REFERENCES


DEFINITIONS, ACRONYMS, ABBREVIATIONS

API:
Application Programming Interface

AR:
Augmented Reality

ASR:
Automatic Speech Recognition

HCI:
Human-Computer Interface (or Interaction)

HRTF:
Head-Related Transfer Function

IP:
Internet Protocol

PC:
Personal Computer

TCP:
Transmission Control Protocol

TTS:
Text-To Speech

UDP:
User Datagram Protocol

WTK:
World Tool Kit