A Distributed Device Diagnostics System Utilizing Augmented Reality and 3D Audio

Reinhold Behringer, Steven Chen, Venkataraman Sundareswaran, Kenneth Wang, and Marius Vassiliou

Rockwell Science Center, Thousand Oaks, CA 91360, USA
phone ++1-805-373-4435, fax ++1-805-373-4862
{reinhold,slchen,vsundar,kkwang,msvassiliou}@rsc.rockwell.com
WWW home page: http://hci.rsc.rockwell.com

Abstract. Augmented Reality brings technology developed for virtual environments into the real world. This approach can be used to provide instructions for routine maintenance and error diagnostics of technical devices. The Rockwell Science Center is developing a system that utilizes Augmented Reality techniques to provide the user with a form of "X-Ray Vision" into real objects. The system can overlay 3D rendered objects, animations, and text annotations onto the video image of a known object, registered to the object during camera motion. This allows the user to localize problems of the device with the actual device in his view. The user can query the status of device components using a speech recognition system. The response is given as an animation of the relevant device module and/or as auditory cues using spatialized 3D audio. The diagnostics system also allows the user to leave spoken annotations attached to device modules for other users to retrieve. The position of the user/camera relative to the device is tracked by a computer-vision-based tracking system especially developed for this purpose. The system is implemented on a distributed network of PCs, utilizing standard commercial off-the-shelf components (COTS).

1 Context and Related Work

During the past few years, technology developed for Virtual Environments (VE) has become a valuable tool for providing an intuitive human-computer interface. In the domain known as *Augmented Reality* (AR), this technology is being applied in the integration of the virtual environment with the real world [11].

1.1 Augmented Reality

Rapid progress in several key areas (wearable computing, virtual reality rendering) has focused significant attention on *Augmented Reality* research in recent years [3]. Although AR is often associated with visualization (starting with the first head-mounted display by Sutherland [20]), augmentation is also possible in the aural [4] [13] and other domains. AR technology provides means of intuitive

information presentation for enhancing situational awareness and perception by exploiting the natural and familiar human interaction modalities with the environment, e.g., augmenting paper drawings [10], a desk environment [16], or familiar collaborative modalities [21].

The concepts of AR have been demonstrated in many applications [1]. AR systems can provide navigational aid in an unknown environment, e.g., in an urban setting [7]. Industrial applications of AR techniques include airplane manufacturing [12], guided assembly [17], improving machine maintenance procedures, and a number of applications in the area of *virtual prototyping*.

1.2 Registration for AR

In applications that utilize the concept of virtual environments, various methods have been developed to track the user's head in order to provide a view which is consistent with the user's sensory information. Such systems include magnetic tracking and attitude sensors. However, visual AR applications require a much higher registration precision, because the human eye is very sensitive to a mismatch between virtual and real objects [1].

Vision-based tracking can potentially provide a very high tracking accuracy. Video-based observer pose estimation methods attempt to compute the position and orientation of the camera from the position of landmarks in the images. The general problem of reliable 3D motion estimation from image features is largely an unsolved problem in computer vision. However, by restricting to the sub-problem of easily identifiable landmarks, the motion estimation problem can be solved (e.g., [9]).

1.3 The RSC Distributed Device Diagnostics System

The Rockwell Science Center (RSC) is developing and integrating components for a system using AR techniques for visualization and auralization during maintenance and error diagnostics procedures. This system is based on an Augmented Reality (AR) approach by overlaying textual information, registered 3D rendered objects, and animations onto a live video of the actual device to be diagnosed, and ultimately by directly overlaying this display into the field of view of the user. 3D audio rendering techniques are used to indicate the location of objects which are currently not in the user's field of view. To employ a non-tethered human-computer interface, the system is operated by speaker-independent speech recognition. To achieve the registration necessary for a well-aligned visual overlay, we have developed a visual tracking module, which relies on tracking of visual fiducial markers and on the mathematical formalism of *visual servoing*. One of the novel features of our concept is the integration of computer vision, speech recognition, AR visualization, and 3D audio in a distributed networked PC environment.

In this paper we describe the system components and the tracking algorithms as well as the status of the system integration.

2 System Components

2.1 Head Tracking: Visual Servoing with Fiducial Markers

The fiducial markers used by the visual tracking system have been designed for easy detectability in clutter and under a wide range of viewing angles. Each marker has a unique ID. Circular markers with concentric rings have a high degree of symmetry, which allows the application of a simple viewpoint-invariant detection algorithm. Similar markers are used by Neumann [14] who developed a color-code scheme for ring marker identification, and Sharma [17] who uses the configuration pattern of a set of markers for detecting ID and orientation of the marker set.

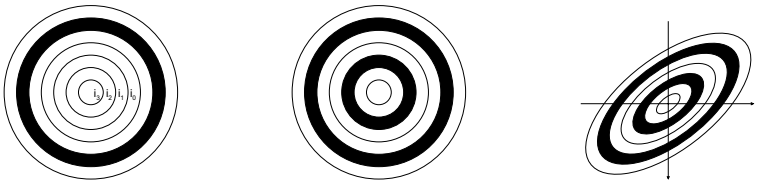

Fig. 1. Left two circles: Schematics of fiducial ring marker. The left marker shows the ring fields i_j which correspond to a binary number. The middle marker, for example, denotes the ID=2. The right marker is distorted to an ellipse when seen under a large viewing angle.

Visual servoing is controlling a system – typically a robot end-effector – based on processing visual information. It is a well-developed theory for robotic vision ([5], [6], [15], [19], [22]). Our application of the visual servoing approach has been outlined in [18], but for completeness, it will briefly be summarized below.

Visual servoing is carried out in a closed-loop fashion, as shown in Fig. 2. We would like the set of system states $\mathbf{s}$ to attain certain target values $\mathbf{s}_r$. The current values of the states $\mathbf{s}$ are measured by a camera looking at the scene. The system uses the error (difference between the target values and current values) to determine the motion parameters T and Ω to move the camera in order to reduce the error. We adopt the standard coordinate systems shown in Fig. 2. The translational velocity T has components U, V, and W. The components of the rotational velocity Ω are A, B and C.

To do this, we need to know the analytical relationship between the motion parameters and the state $\mathbf{s}$. Usually, the forward relationship, namely the change in $\mathbf{s}$ due to parameters T and Ω is known. The goal is to minimize $\|\mathbf{s} - \mathbf{s}_r\|$. Let us define the error function

$$\mathbf{e} = \mathbf{s} - \mathbf{s}_r \tag{1}$$

We would like the error function to decay exponentially:

$$\dot{\mathbf{e}} = -\lambda \cdot \mathbf{e},$$

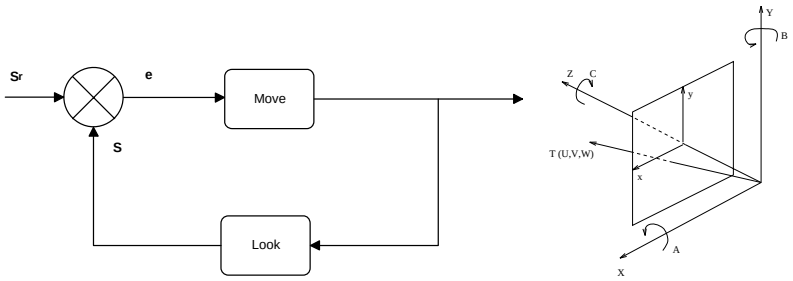

Fig. 2. Schematic of the visual servoing approach (left) and the coordinate system (right).

where λ, the constant in the exponential, controls the decay rate (i.e., speed of convergence). Therefore $\dot{\mathbf{s}} = -\lambda \cdot (\mathbf{s} - \mathbf{s}_r)$. From standard optic flow equations (see for e.g. [8]), we know that we can write the 2D displacement of an image feature at (x_p, y_p) as:

$$\begin{aligned} \dot{x_p} &= \tfrac{1}{Z(x_p,y_p)} \left[-U + x_p W\right] + A x_p y_p - B\left[1 + x_p^2\right] + C y_p, \\ \dot{y_p} &= \tfrac{1}{Z(x_p,y_p)} \left[-V + y_p W\right] + A\left[1 + y_p^2\right] - B x_p y_p - C x_p. \end{aligned} \tag{2}$$

We assume that the images are planar, obtained by the pin-hole perspective approximation with a focal length of unity (see Fig. 2). This relationship between change in 2D projection of a point and the motion parameters is of the form

$$\dot{\mathbf{s}} = L \begin{pmatrix} T \\ \Omega \end{pmatrix}, \tag{3}$$

where L is the "interaction matrix" consisting of 2D coordinates (x_p, y_p) and the depth Z of the 3D point projected at (x_p, y_p), T is the translation vector and Ω is the rotation vector. We would like to determine T and Ω. Assuming that the motion of features $\mathbf{s}$ is due to the motion T and Ω, we obtain:

$$L \begin{pmatrix} T \\ \Omega \end{pmatrix} = -\lambda \mathbf{e}. \tag{4}$$

Inverting Eqn. 4, we get the control law

$$\begin{pmatrix} T \\ \Omega \end{pmatrix} = -\lambda L^{+} \mathbf{e}, \tag{5}$$

where L^{+} is the pseudo-inverse of L.

This allows us to compute the motion of the camera required to minimize the error $\mathbf{e}$. When performed in closed-loop, the value $\mathbf{s}$ will reach $\mathbf{s}_r$ when error $\mathbf{e}$ is reduced to zero.

2.2 The 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. It 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^{TM}. Using the ViaVoice engine, speaker-independent continuous phonetic recognition constrained by finite state grammars is possible, as well as speaker-adapted continuous dictation using an American English language model. 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 application may be running on MS-Windows, Solaris, IRIX, or any other operating system that supports TCP/IP networking. Using a serial-like ASCII command protocol, the client application indicates its identity to the server, as well as any contextual data of relevance to the speech recognition task, such as the currently selected object in the graphical portion of the client application's user interface. Speech recognition is requested and activated by the client, and asynchronous speech recognition results are sent from the ASR Server to the client. Both of these interactions occur using a vendor-independent protocol.

Speech is currently acquired locally through a sound card installed in the PC on which the ASR Server executes, although an add-on option for uncompressed streaming audio over IP networks has been developed to allow the microphone to reside on another computer. Recognition results are reported to the client application immediately (per word) and/or upon the completion of a whole utterance (sentence).

If the client application's user interface can wait for whole utterance results, an application-specific text parser in the ASR Server may be exploited to relieve the client application of the burden of parsing the recognition results. This provides an advantage in the application design phase when rapidly iterating through grammar and vocabulary designs. Per-word confidence scores can be reported to the client application if requested; in addition, reporting of word timing hypotheses is being developed – this capability will enable concurrent gesture and speech interfaces.

In the device diagnostics system, described in this paper, speech is used both to control the operation (initialization, mode switching), and to ask questions about device modules, such as the query "Where is the Power Supply?". Moreover, the dictation recognition mode is exploited to attach textual "virtual notes" to selected objects in the virtual environment. The recognition is started by a trigger button pressed by the user.

2.3 AR Visualization

The AR visualization provides visual rendering of the device which is to be diagnosed, and its components. The following items can be overlaid in real-time in the live video image:

- A CAD wireframe model of the outer device shape.
- CAD models of the interior components of the device.
- Textual annotations, attached to components of the device.

The CAD models are overlaid onto the video image of the device and, therefore, provide a kind of "X-ray vision" into the interior of the device. They can be animated to blink between *shaded* and *wireframe* display rendering in order to highlight the corresponding module of the device. In Fig. 3 (right) the two rendering modes are shown. The rendering engine used for creating the 3D overlay is the Sense8 World-Toolkit library.

2.4 3D Audio Auralization

A three-dimensional (3D) audio system provides an auditory experience in which sounds appear to emanate from locations in 3D space. 3D audio can be used to indicate spatial locations as well as increase the differentiability of multiple audio communication channels. Thus, both visual display clutter and message comprehension time may be reduced through the use of 3D audio. Head-Related Transfer Functions (HRTF) [2] are basically filters incorporating the effects of the sound signal reflecting off the listener's head, shoulders, and pinnae (outer ears). These HRTFs are usually different for each listener. However, it has been shown that the HRTFs of a "good localizer" are suitable for a large group of users [23]. A good localizer is defined as a human who localizes real sound sources with high accuracy and precision.

In our device diagnostics system, we use the commercial Aureal Semiconductor off-the-shelf (COTS) 3D audio system. This comprises a software application programming interface (API) based on the Microsoft DirectSound/DirectSound3D standard as well as a PC sound card including a chip designed and manufactured by Aureal. The inputs of an HRTF-based 3D audio system include the monaural sound source signal(s) (one signal per sound source), the user position and orientation, and the positions of the sound source(s). The 3D audio is output over a pair of headphones or two speakers. In order to provide 3D audio capability to non-platform-specific applications, a TCP/IP sockets server (the RSC 3DA Server) was developed. This allows application developers to simply exploit the Aureal 3D audio services by establishing a socket connection to the RSC 3DA Server and providing real-time user position and orientation and sound source position data. The sound source signals are stored as wave files on the 3DA Server host PC. The 3DA Server operates at 30 fps, and current COTS 3D audio sound cards support up to three 44.1 kHz-sampled sound sources.

3 System Integration

The system components of the AR device diagnosis system were implemented on a PC. We also used a standard tower PC as the example *device* which was to be "diagnosed". A CAD model was hand-coded to describe the outer geometry and a few inner components of the PC: CD ROM drive, network card, power supply, and structural components. The AR visualization is implemented on a 200 Mhz PC, running under Windows NT. The rendering algorithms are based on the Sense8 WorldToolkit library, which provides functionality for the display of 3D worlds. The implementing PC is equipped with a Imaging Technology color framegrabber, which digitizes the video signal from a Cohu 2200 CCD camera. In order to illustrate potential diagnosis capabilities, we introduced the following simulated device errors in the PC being "diagnosed": CD ROM failure, power supply overheating, and network card failure. The spatial interpretation for the 3D auralization is obtained solely from the visual servoing algorithm.

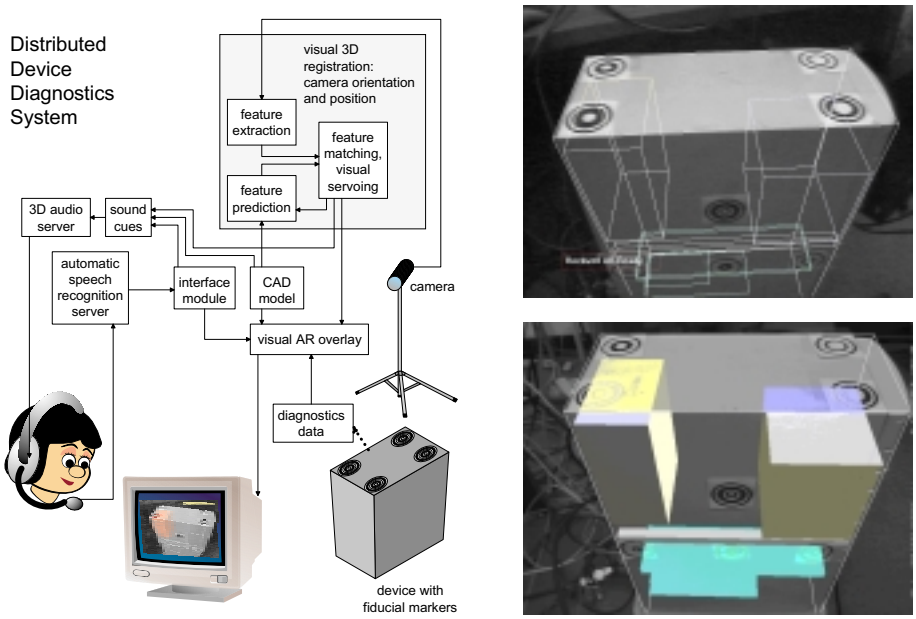

Fig. 3. Left: Concept of the current implementation of the device diagnostics system. Right: Registered wireframe and shaded overlay onto the video image.

The speech recognition server (ASR server) and the 3D audio server are both running on another PC (166 Mhz) under Windows 95. The user can query by voice command the location of the CD ROM drive, network card, and power supply. A flashing animation, overlaid on the video, visualizes the location. The user can also query the location of printer and UPS. Their location is indicated by a 3D audio cue: spatialized sounds "move" in the direction of the location and guide the user.

4 Experimental Results

The visual AR overlay is rendered with a framerate between 6-10 fps, depending on the system load. The framerate is slowed down by the image transfer implementation in Windows, which is currently not optimized for speed.

Fig. 4 to Fig. 6 show various display modalities of the visual overlay on the video image. Fig. 4 shows the wireframe overlay from two different directions. The overlay matches very well with the real object under a wide viewing angle range. Not surprisingly, the registration error depends on the number (and the location) of the recognized fiducial markers. The alignment/registration is slightly off for more extreme viewing angles where only markers in one plane are well visible. The markers in the other plane are too slanted for robust recognition. However, the range for acceptable registration is quite large, and the error is acceptable.

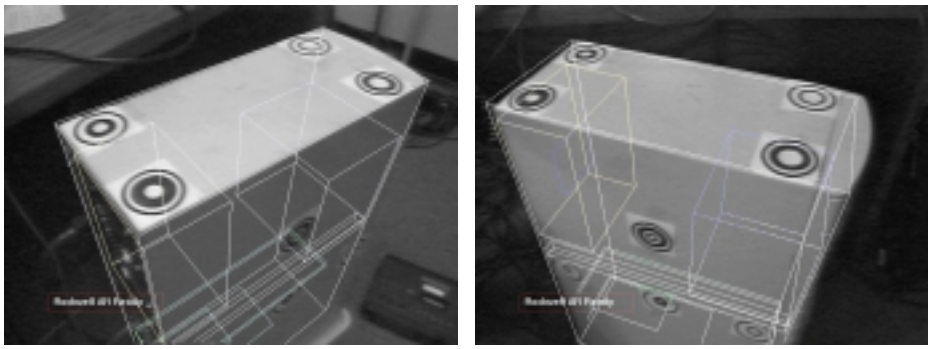

Fig. 4. Overlay of wireframe onto video stream from different directions.

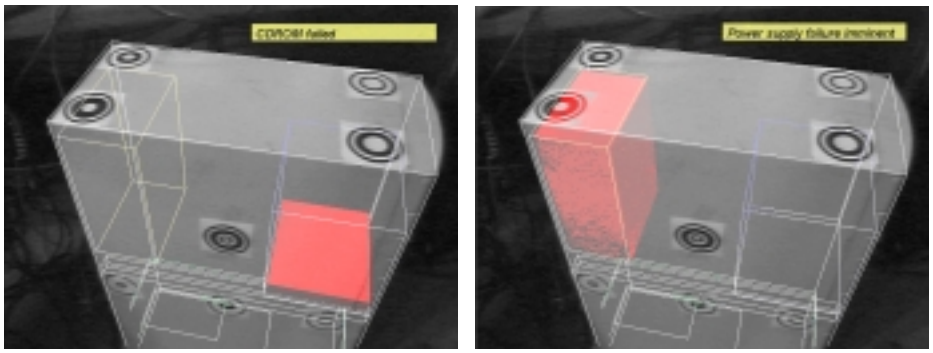

Fig. 5. Overlay of error indication onto video image: CD ROM and power supply.

In Fig. 5 the error indication is shown by a flashing volumetric rendering of the relevant PC components. In Fig. 6 two examples of annotations are shown.

These notes have been left by a previous user who attached these notes to components of the PC. They are spoken and translated by the ASR text-to-speech (TTS) system into ASCII text.

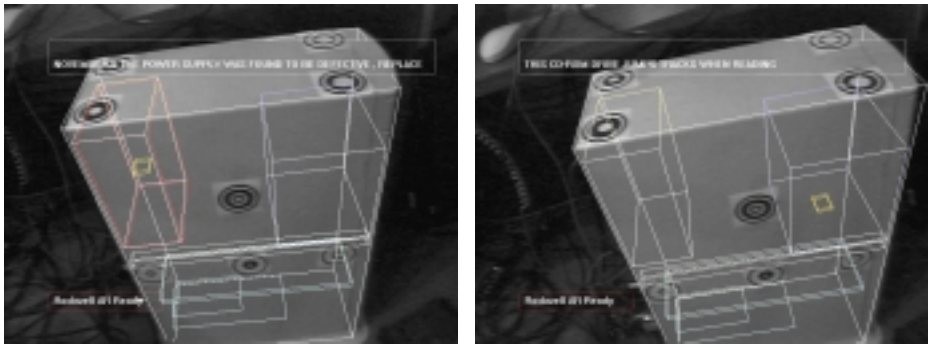

Fig. 6. Retrieval of textual annotations.

5 Summary and Conclusion

We have demonstrated the capability of a distributed, networked system for device diagnostics as a novel, tetherless interface for human-computer interaction. The distributed architecture of the system ultimately allows for scalability and enables the user to be equipped with only a light, wearable computer system that provides wearable display capabilities. Such a system can one day replace service manuals by providing direct overlay of maintenance instructions or error diagnosis onto the real object.

References

[1] Azuma, R. T. A survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments 6*, 4 (1997), 355–385.

[2] Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1997.

[3] Caudell, T. P. Introduction to Augmented and Virtual Reality. In *Proc. of SPIE Conf. on Telemanipulator and Telepresence Technologies* (Boston, MA, Oct. 1994), SPIE, pp. 272–281.

[4] Cohen, M., Aoki, S., and Koizumi, N. Augmented audio reality: Telepresence/VR hybrid acoustic environments. In *Proc. of Workshop on Robot and Human Communication* (Tokyo, Japan, Nov. 1993), IEEE Press, pp. 361–4.

[5] Espiah, B., Chaumette, F., and Rives, P. A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation 8*, 3 (1992), 313–326.

[6] Feddema, J., and Mitchell, O. Vision-guided servoing with feature-based trajectory generation. *IEEE Transactions on Robotics and Automation 5*, 5 (1989), 691–700.

[7] Feiner, S., MacIntyre, B., Höllerer, T., and Webster, A. A touring machine: prototyping 3D mobile Augmented Reality systems for exploring the urban environment. In *Proc. of 1st In. Symp. on Wearable Computers* (Cambridge, MA, Oct. 1997), pp. 74–81.

[8] Horn, B. K. P. *Robot Vision*. MIT Press, Cambridge, 1987.

[9] Koller, D., Klinker, G., Rose, E., Breen, D., Whitaker, R., and Tuceryan, M. Real-time vision-based camera tracking for Augmented Reality applications. In *Proc. of VRST '97* (Lausanne, Switzerland, Sept. 1997).

[10] Mackay, W., Pagani, D., Faber, L., Inwood, B., Launiainen, P., Brenta, L., and Pouzol, V. ARIEL: Augmenting paper engineering drawings. In *Proc. of Conf. on Human Factors in Computing Systems (CHI)* (Denver, CO, May 1995), IEEE Press, pp. 421–422.

[11] Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. Augmented Reality: a class of displays on the reality-virtuality continuum. In *Proc. of SPIE Conf. on Telemanipulator and Telepresence Technologies* (Boston, MA, Oct. 1994), SPIE, pp. 282–292.

[12] Mizell, D. Virtual reality and Augmented Reality in aircraft design and manufacturing. In *Proc. of Wescon Conference* (Anaheim, CA, Sept. 1994), p. 91ff.

[13] Mynatt, E. D., Back, M., Want, R., and Frederick, R. Audio Aura: lightweight audio Augmented Reality. In *Proc. of ACM UIST '97* (Banff, Canada, Oct. 1997), ACM, pp. 211–12.

[14] Neumann, U., and Cho, Y. Multi-ring fiducial systems for scalable fiducial Augmented Reality. In *Proc. of VRAIS '98* (Atlanta, Mar. 1998).

[15] Papanikolopolous, N., Khosla, P., and Kanade, T. Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision. *IEEE Transactions on Robotics and Aautomation 9*, 1 (1993), 14–35.

[16] Rauterberg, M., Mauch, T., and Stebler, R. Digital playing desk: A case study for Augmented Reality. In *Proc. of IEEE Workshop on Robot and Human Communication* (Tsukuba, Japan, Nov. 1996), IEEE Press, pp. 410–415.

[17] Sharma, R., and Molineros, J. Computer vision-based Augmented Reality for guiding manual assembly. *PRESENCE: Teleoperators and Virtual Environments 6*, 3 (June 1997), 292–317.

[18] Sundareswaran, V., and Behringer, R. Visual servoing-based Augmented Reality. In *Proc. of First Int. Workshop on Augmented Reality (IWAR) '98* (San Francisco, CA, Nov. 1998).

[19] Sundareswaran, V., Bouthemy, P., and Chaumette, F. Exploiting image motion for active vision in a visual servoing framework. *International Journal of Robotics Research 15*, 6 (1996), 629–645.

[20] Sutherland, I. E. A head-mounted three dimensional display. In *Proc. of Fall Joint Computer Conference* (Washington, DC, 1968), Thompson Books, pp. 757–764.

[21] Szalavari, Z., Gervautz, M., Fuhrmann, A., and Schmalstieg, D. Augmented Reality enabled colaborative work in "Studierstube". In *Proc. of EUROVR '97* (Amsterdam, The Netherlands, 1997).

[22] Weiss, L., Sanderson, A., and Neumann, C. Dynamic sensor-based control of robots with visual feedback. *IEEE Transactions on Robotics and Automation 3*, 5 (1987), 404–417.

[23] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. Localization using non-individualized head-related transfer functions. *Journal of the Acoustical Society of America* (1993), 111–123.