A Novel Interface for Device Diagnostics Using Speech Recognition, Augmented Reality Visualization, and 3D Audio Auralization

Reinhold Behringer, Steven Chen, Venkataraman Sundareswaran, Kenneth Wang, and Marius Vassiliou
Rockwell Science Center, Thousand Oaks, CA 91360, USA
{reinhold,slchen,vsundar,kkwang,msvassiliou}@rsc.rockwell.com
WWW home page: http://hci.rsc.rockwell.com

Abstract

Routine maintenance and error diagnostics of technical devices can be greatly enhanced by applying multimedia technology. The Rockwell Science Center is developing a system which can indicate maintenance instructions or diagnosis results for a device directly into the view of the user by utilizing Augmented Reality and multimedia techniques. The system can overlay 3D rendered objects, animations, and text annotations onto the live video image of a known object, captured by a movable camera. The status of device components can be queried by the user through a speech recognition system. The response is given as an animation of the relevant device module, overlaid onto the real object into the user's view, and/or as auditory cues using spatialized 3D audio. The position of the user/camera relative to the device is tracked by a computer vision based tracking system. The diagnostics system also allows the user to leave spoken annotations attached to device modules for other users to retrieve. The system is implemented on a distributed network of PCs, utilizing standard commercial off-the-shelf (COTS) components.

1. Introduction and context

Multimedia technology is often used in the context of entertainment applications. However, the capabilities of this technology can also be applied to "serious" industrial applications such as like maintenance and diagnostics of an industrial device. The integration of visual and auditory displays can provide a more in-depth understanding of the maintenance procedures and the device functionality and therefore can enhance the awareness of the person in charge of device repairs.

The Rockwell Science Center (RSC) is developing and integrating components for a system using multimedia 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, 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 techniques are used to indicate cues about objects which are currently not in the user's field of view. In order to avoid a tethered human-computer interface, the system is operated by speaker-independent speech recognition. For achieving registration which is necessary for a well-aligned visual overlay, we have developed a visual tracking module, relying on tracking of visual fiducial markers. The novelty 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 current status of the integration of the system.

1.1. Augmented reality

Progress in the research areas of *wearable computing* and *virtual reality rendering* in recent years has enabled rapid development of *Augmented Reality* (AR) technology. This 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. Although AR is often associated with visualization (starting with the first head-mounted display by Sutherland [24]), augmentation is also possible in the aural domain [7] [17].

The concepts of AR have been applied in many applications. AR systems can provide navigational aid in an unknown environment, e.g., in an urban setting [10], or serve as a tour guide in a museum [3]. AR systems have been proposed and demonstrated for architecture [26], interior design [1], and related fields. Important industrial AR applications within the scope of this paper are machine main-

tenance, airplane manufacturing [16], and guided assembly [20]. In medicine, AR technology has been reported to be applied as an important auxiliary tool (e.g., [4], [6]). Its primary purpose here is to visualize data, acquired from ultrasound [21], MRI and CT [12] or other sources, and overlay them onto the patient's body [25]. Completely immersive AR requires the user to wear head-mounted see-through displays which require precise calibration [15]. However, a different way of applying AR techniques is to overlay information onto a video (stream or single snapshot). Such systems do not provide user immersion, unless the video is captured by a head-mounted camera and is displayed through a head-mounted display to the user. Nevertheless, such video AR can also be useful in providing more insight into processes, objects, or environment.

1.2. Registration for AR

A very important issue in AR is the registration between the information to be displayed and the real world: the human eye can detect differences smaller than 1 minute of arc. Currently there is no system which can provide an accuracy high enough for completely merging the virtual world with the real environment. However, in order to obtain more or less satisfying registration, a precision of 0.5 deg seems to be adequate [2].

Indoor AR applications can employ a variety of tracking systems, which were originally developed for VR applications, e.g. magnetic trackers. Magnetic tracking can readily provide the position and orientation of the observer (e.g., [21]). The major limitations of magnetic tracking are its short range (typically 8 ft radius) and sensitivity to metallic objects in the vicinity. The absolute orientation angles can also be obtained by tilt and roll sensors. Their accuracy for static and quasi-static measurements is acceptable, but they have a long lag. Inertial sensors can provide data during rapid motion, but they tend to induce drift. Hybrid tracking systems are a promising alternative. They combine inertial tracking with static position measurements. Such systems are already commercially available (e.g., the Intersense inertial tracking system with ultrasound positioning [11]).

The registration method which promises the highest accuracy, but also has the highest hurdles, is computer vision-based. Many AR systems employ computer vision technology for achieving the required registration precision. The approach in most cases is to detect visual features which have a known position, and recover camera orientation and position through a matching or pose recovery process. Object pose estimation methods determine the position and orientation of the object, for e.g., a plane containing landmarks. Typically, the landmarks are located using image processing, and the pose is determined in each frame. The pose is then used to render the 3D model. 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. Our approach is based on the mathematical formalism of *visual servoing*.

2. HCI components

The components of the diagnosis system are a tracking system using visual fiducial markers, a speech recognition system for user input and system control, an AR visualization system producing a 3D overlay, and a 3D auditory display providing clues for regions which are not in the user's field of view.

2.1. Tracking and registration

To provide a registered AR overlay, the camera must be tracked to obtain position and orientation relative to the device to be diagnosed. If only approximate registration is required, the tracking can be performed by conventional methods (e.g., magnetic tracking). However, these methods have been proven to provide insufficient registration precision for AR purposes. Therefore, a new approach, using visual fiducial markers for fast high precision tracking, has been developed.

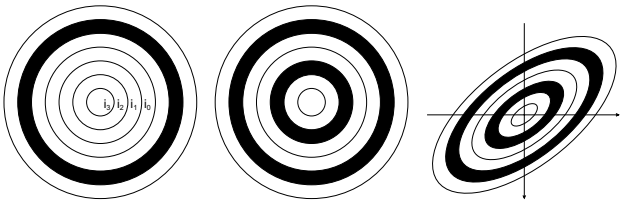

Figure 1. Schematics of the fiducial markers.

To be useful for registration, fiducial markers must be clearly detectable even in a cluttered environment and should be distinguishable from each other by a unique identification. Marker detection and identification must be reliable from a wide range of viewing angles. These conditions can be fulfilled by the use of circular fiducial markers. Such markers, also used by Neumann [18] and Hoff [13], have a high degree of symmetry, which allows the application of a simple viewpoint-invariant detection algorithm. To provide markers with a unique ID tag, Neumann developed a color-code scheme for ring marker identification. Our approach uses a marking scheme, based on a pattern of concentric rings.

In Fig. 1 the marking scheme is shown. The fiducial markers are identified by their outer black ring. This concentric ring marking scheme has the advantage of being scale-invariant (at least within the boundaries given by camera field-of-view and pixel resolution). The diameter of the outer ring provides a norm for reading out the inner marker pattern. When seen from an arbitrary viewing angle, the circular ring pattern is seen as a concentric ellipse pattern.

In Fig. 2 the detection and identification of 4-bit ring markers is shown in a realistic lab scenario with a cluttered background (cables on the right side). The system is able to detect and identify all 16 markers on the test sheet correctly. It also detects a ring candidate within the cable clutter, but it rejects it due to non-conformance with the constraint of smooth shape (marked by the system as ID=-2). Searching through a complete image of 640×480 pixels, as shown in Fig. 2, requires about 0.4 sec time on a 200 MHz Pentium Pro. Tracking the rings in smaller search windows requires much less computation time (40-80 ms).

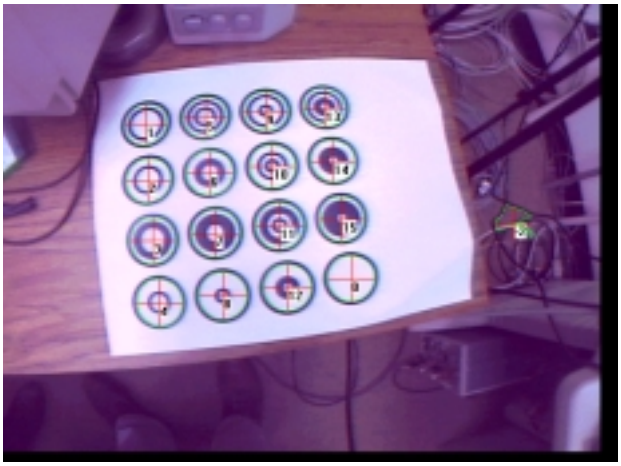

Figure 2. Example of ring detection and identification for 4-bit markers.

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 (see e.g., [8], [9], [19], [23] [27]). Visual servoing is carried out in a closed-loop fashion, as shown in Fig. 3. Its application to motion estimation for AR registration has been described in [22], but it shall briefly be summarized in this paper for completeness. 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. The translational velocity T has components U, V, and W. The components of the rotational velocity Ω are A, B and C.

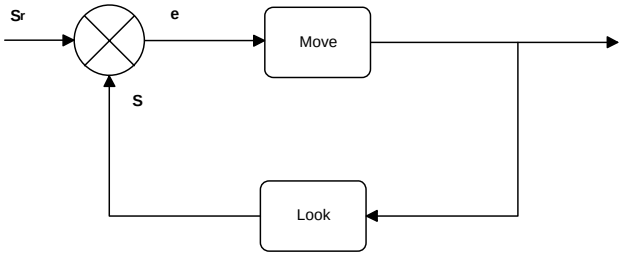

Figure 3. Schematics of visual servoing.

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 $\|\mathbf{e}\| = \|\mathbf{s} - \mathbf{s}_r\|$. From standard optic flow equations (see for e.g. [14]), we know that we can write the 2D displacement of an image feature at (x_p, y_p) as:

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

We assume that the images are planar, obtained by the pinhole perspective approximation with a focal length of unity (see Fig. 3). This relationship between change in 2D projection of a point and the motion parameters is of the form

$$\dot{\mathbf{s}} = L \begin{pmatrix} T \\ \Omega \end{pmatrix}, \tag{2}$$

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:

$$L \begin{pmatrix} T \\ \Omega \end{pmatrix} = -\lambda \mathbf{e}. \tag{3}$$

Inverting Eqn. 3, we get the control law

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

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 speech recognition 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. 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.

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.

Recognition results are reported to the client application immediately (per word) and/or upon the completion of a whole utterance (sentence). Per-word confidence scores can be reported to the client application if requested; in addition, reporting of word timing hypotheses is currently under development – this capability will enable concurrent gesture and speech interfaces. In the device diagnostics system, speech input is used to switch between modes of operation and to ask questions about objects, such as the query, "Where is the Power Supply," when inspecting the wireframe model of a personal computer. Moreover, the dictation recognition mode can be exploited to attach textual "virtual notes" to selected objects in the virtual environment.

2.3. Visualization by AR techniques

The AR visualization module can render the following overlay items onto a live video image: A CAD wireframe model of the outer device shape, CAD models of the interior components of the device, and textual annotation which the user can speak into the microphone and "attach" to a device component. The wireframe overlay provides a strong visual clue for manually judging the registration accuracy. This AR overlay concept provides a kind of "X-ray vision" into the interior of the device. The rendering of device components blinks between *shaded* and *wireframe*, in order to highlight critical areas. The 3D rendering is done based on the results from the head tracking module in order to align the rendered world with the real world.

2.4. Auralization by 3D audio

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 to 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. In the real world, humans utilize three basic spatial hearing cues: the Interaural Time Difference (ITD), the Interaural Intensity Difference (IID), and the Head-Related Transfer Function (HRTF) [5]. The former two are simply the sound signal phase and amplitude differences respectively between the left and right ears of the listener. The latter is a filter incorporating the effects of the sound signal reflecting off the listener's head, shoulders, and pinnae (outer ears). At RSC, the application of HRTF-based 3D audio is being used as the main cue to create spatialized auditory cues. Ideally, the personalized HRTFs of a listener would be utilized. However, it has been shown that HRTFs recorded from a "good localizer" – a person who has the ability or localizing sound sorces with high precision – provide good 3D audio clues for a large number of listeners [28].

In the RSC device diagnostics system, the Aureal Semiconductor commercial, off-the-shelf (COTS) 3D audio system is being used. This system 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. 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 RSC 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 error diagnosis system were implemented by using a standard tower PC as the *device* which was to be diagnosed. A CAD model was hand-coded, describing 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

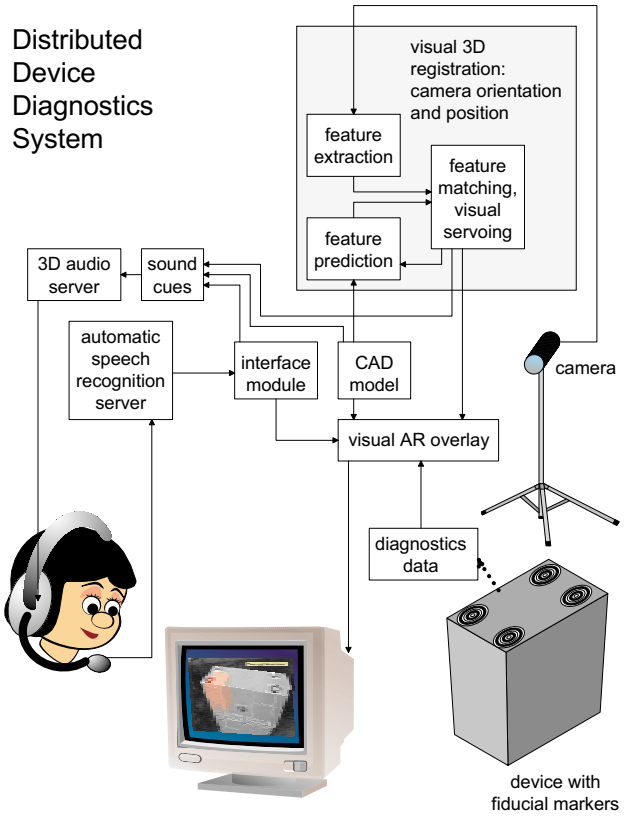

Figure 4. Architecture of the system.

WorldToolkit library, which provides functionality for the display of 3D worlds. The PC is equipped with a Imaging Technology color framegrabber, which digitizes the video signal from a Cohu 2200 CCD camera. The visualization is currently implemented by video overlay. In a later stage, we intend to use a head-mounted see-through display for direct overlay into the observer's view. The simulated errors are CD ROM failure, power supply overheating, and network card failure. At the current stage, the camera plays the role of the observer. Besides the visual tracking, there is no additional tracking system implemented. The spatial interpretation for the 3D auralization is obtained solely from the visual servoing algorithm.

The speech recognition server (ASR server) and the 3D audio server are both running on another PC (166 Mhz) under Windows 95. This PC is connected to a local ethernet network over IP. The microphone is connected directly to this PC. The user can query the location of 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. Since these devices are not in the field of view, 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 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. In Fig. 5 the overlay is shown as wireframe and with shaded components. It can be seen that the alignment/registration precision is very high. Experiments indicated that the overlay 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. The range for acceptable registration is about ± 50 deg.

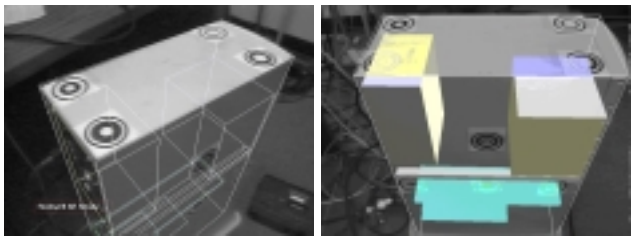

Figure 5. Overlay of wireframe (left) and volumetric shaded (right) interior components.

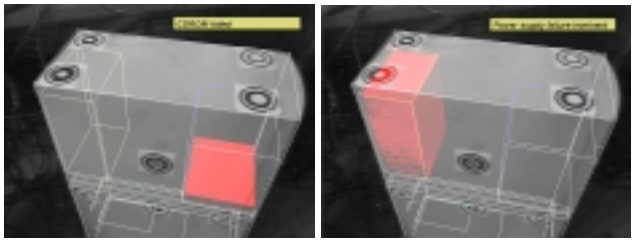

Figure 6. Overlay of error indication.

In Fig. 6 the error indication is shown by a volumetric rendering of shaded PC components. The volumetric rendering actually flashes to highlight the problem zone.

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 system uses multimodal cues (visual, aural) to attract user attention, and employs advanced AR techniques to improve the user awareness of the status of the technical device to be monitored, here demonstrated with a PC.

The distributed architecture of the system ultimately allows for scalability and enables the user to be equipped with only a light, wearable computer system which provides wearable display capabilities. Such a system can replace

service manuals by providing direct overlay of maintenance instructions or error diagnosis onto the real object.

Progress is necessary in providing high network bandwidth for effectively reducing the computation power of the computer system worn by the user. More attention must also be focused on the tracking modules. The goal is to employ a hybrid tracking system which utilizes tracking methods that compensate for each other's shortcomings.

References

[1] K. Ahlers, A. Kramer, D. Breen, P.-Y. Chevalier, C. Crampton, E. Rose, M. Tuceryan, R. Whitaker, and D. Greer. Distributed Augmented Reality for collaborative design applications. In *Proc. Eurographics '95*, pages C13–C14, Maastricht, The Netherlands, 1995.

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

[3] B. B. Bederson. Audio Augmented Reality: a prototype automated tour guide. In *Proc. of Conf. on Human Factors in Computing Systems (CHI)*, pages 210–211, Denver, CO, May 1995.

[4] J. W. Berger, M. E. Leventon, N. Hata, W. M. W. III, and R. Kikinis. Design considerations for a computer-vision-enabled opthalmic Augmented Reality environment. In *Proc. of CVRMED/MRCAS*, Grenoble, France, 1997.

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

[6] P. Brodeur, J. Dansereau, J. de Guise, and H. Labelle. Points-to-surfaces matching technique for the application of Augmented Reality during spine surgery. In *Proc. of IEEE Conf. on Engineering in Medicine and Biology*, pages 1197–1198, Montreal, Canada, Sept. 1995.

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

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

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

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

[11] E. Foxlin, M. Harrington, and Y. Altshuler. Miniature 6-DOF inertial system for tracking HMDs. In *Proc. Aerosense '98*, pages 48–51, Orlando, FL, Apr. 1998.

[12] W. E. L. Grimson, G. J. Ettinger, S. J. White, T. Lozano-Perez, W. M. W. III, and R. Kikinis. An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality visualization. *IEEE Transactions on Medical Imaging*, 15(2):129–140, Apr. 1996.

[13] W. A. Hoff and K. Ngyen. Computer vision-based registration techniques for Augmented Reality. In *Proc. of Intelligent Rootics and Computer Vision XV*, volume 2904 of *SPIE Intelligent Systems and Advanced Manufacturing*, pages 538–48, Boston, MA, Nov. 1996. SPIE.

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

[15] A. L. Janin, D. Mizell, and T. Caudell. Calibration of head-mounted displays for Augmented Reality applications. In *Proc. of VRAIS '93*, pages 246–55, Seattle, WA, Sept. 1993.

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

[17] E. D. Mynatt, M. Back, R. Want, and R. Frederick. Audio Aura: light-weight audio Augmented Reality. In *Proc. of ACM UIST '97*, pages 211–12, Banff, Canada, Oct. 1997. ACM.

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

[19] N. Papanikolopolous, P. Khosla, and T. Kanade. 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):14–35, 1993.

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

[21] A. State, M. A. Livingston, W. F. Garrett, G. Hirota, M. C. Whitton, E. D. Pisano, and H. Fuchs. Technologies for Augmented Reality systems: Realizing ultrasound-guided needle biopsies. In *Proc. of SIGGRAPH*, New Orleans, LA, 1996. ACM Press.

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

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

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

[25] M. Uenohara and T. Kanade. Vision-based object registration for real-time image overlay. In *Proc. of 1st Int. Conf. on Computer Vision, Virtual Reality, and Robotics in Medicine*, Nice, France, Apr. 1995.

[26] A. Webster, S. Feiner, B. MacIntyre, W. Massie, and T. Krueger. Augmented Reality in architectural construction, inspection, and renovation. In *Computing in Civil Engineering*, pages 913–919, New York, NY, 1996.

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

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