

# **PARIS: FUSING VISION-BASED LOCATION TRACKING WITH STANDARDS-BASED 3D VISUALIZATION AND SPEECH INTERACTION ON A PDA**

*Stuart Goose, Sinem Güvenç, Xiang Zhang, Sandra Sudarsky, Nassir Navab*

Multimedia Technology Department, Siemens Corporate Research, Inc.

755 College Road East, Princeton, NJ 08540, USA

†Columbia University, 500 West 120th Street, New York, NY, 10027, USA

## **ABSTRACT**

Industrial service and maintenance is by necessity a mobile activity, and the aim of the technology reported is towards improving automated support for the technician in this endeavor. As such, a framework was developed called PARIS (PDA-based Augmented Reality Integrating Speech) that executes entirely on a commercially available PDA equipped with a small camera and wireless support. Real-time computer vision-based techniques are employed for automatically localizing the technician within the plant. Once localized, PARIS offers the technician a seamless multi-modal user interface juxtaposing a VRML augmented reality view of the industrial equipment in the immediate vicinity and initiates a context-sensitive VoiceXML speech dialog concerning the equipment. Integration with the plant management software enables PARIS to access equipment status wirelessly in real-time and present it to the technician accordingly.

## **1. INTRODUCTION AND MOTIVATION**

Siemens is the world's largest supplier of products, systems, solutions and services in the industrial and building technology sectors. Service and maintenance is by necessity a peripatetic activity, and as such one continuing aspect of our research focuses upon improving automated support for this task. Another future trend that we have been focusing on is applying 3D interaction and visualization techniques to the industrial automation domain.

In recent years we have witnessed the remarkable commercial success of small screen devices, such as cellular phones and Personal Digital Assistants (PDAs). Keyboards remain the most popular input device for desktop computers. However, performing input efficiently on a small mobile device is more challenging. Speech interaction on mobile devices has gained in currency over recent years, to the point now where a significant proportion of mobile devices support or include some form of speech recognition.

The ability to model real world environments and augment them with animations and interactivity has benefits over conventional interfaces. However, navigation and manipulation in 3D graphical

environments can be difficult, and disorientating, especially when using a conventional mouse. Small sensors can be used to report various data about the surrounding environment and relative movement, etc. One such sensor is that of a small camera.

The hypothesis that motivated this research is that a camera, in conjunction with computer vision algorithms, could be exploited to provide location information, which in turn, could seamlessly and automatically drive the navigation through a 3D graphical world representing selected elements in the real world. In addition to eradicating partially the complexity of 3D navigation, integrating context-sensitive speech interaction could further simplify and enrich the mobile interaction experience. Hence, the PARIS framework was developed for experimenting with the provision of mobile, context-sensitive, multi-modal user interfaces for mobile maintenance.



**Figure 1: A mobile maintenance technician using PARIS.**

To the knowledge of the authors, this is the first reported VRML-based AR framework that executes entirely on a commercially available PDA. PARIS employs real-time vision algorithms for localizing a technician and offers a multimodal user interface that synchronizes an augmented reality graphical view based on VRML [22] with a VoiceXML [21] speech-driven interface. After automatically detecting when the technician enters the vicinity of a specific plant component, PARIS can engage him or her in a context-

specific speech dialog concerning the corresponding component, as shown in figure 1.

Reported in the remainder of the paper are some novel aspects of PARIS. A brief discussion of related work is provided in Section 2. The system architecture and components are presented in section 3. In section 4, the VoiceXML Lite subsystem that provides location and context-sensitive speech support is reported. Our vision-based techniques for marker tracking and localization are presented in section 5. The industrial scenario and multimodal interaction experience is offered in section 6. Sections 7 and 8 propose areas for further research and provide some concluding remarks.

## 2. RELATED WORK

This review selectively traces the progress of mobile 3D interfaces, location tracking, augmented reality and speech interaction, and hence the confluence of these technologies as they relate to mobile maintenance.

The benefits of mobile maintenance [20] and virtual environments [4] to the industrial sector have been reported. Navigation and manipulation in desktop 3D graphical environments can be difficult. This need spawned research into novel input and control devices for this purpose [25]. Fitzmaurice *et al* [6] in 1993 simulated a palmtop computer to, among other things, evaluate how novel input devices can expedite interaction in virtual environments on handheld devices. Hinckley *et al* [12] describes how a PocketPC was augmented with multiple sensors to offer adaptive interaction with mobile devices, including automatic power on/off, automatic landscape/portrait flipping etc. Mobile Reality [10] is a framework that combines the input from infrared beacons and an inertia tracker to drive automatically the VRML display on a PDA. In contrast to the above, PARIS leverages vision-based localization algorithms executing on the PDA to adjust the viewpoints in the VRML scene correspondingly.

A variety of indoor location tracking technologies have been reported. The Active Badge System [24] facilitates position tracking of people wearing badges in an office environment and, for example, to route phone calls to the closest telephone. Memoclip [2] aims at providing users with location-based messages. When a user approaches a sensor that is physically associated with a reminder, the Memoclip displays the corresponding message. Newman *et al* [15] describe an AR system whereby the user wears an ultrasonic positioning device. An X-Windows server redirects the user interface of the application to an iPAQ running Linux. Goose *et al* [10] report a PDA-based hybrid tracking solution that fuses the input from infrared beacons and a three degrees-of-freedom (3 DOF) inertia tracker.

The German Ministry for research and training (BMBF) funds a project called AR-PDA [1]. A few

research groups have focused their attention on the use of PDAs for augmented reality applications [3, 7, 8, 9], however none of these perform image processing onboard the PDA. Bertelsen *et al* [3] use the PDA in conjunction with a barcode reader to access to the data in a water treatment plant. Geiger *et al* [7, 9] use the PDA to acquire images, wirelessly transmit them to a server for marker detection, scene augmentation and retransmission back to the PDA for AR visualization. Gausemeier *et al* [9] proceeds in a similar vein but try a method that exploits feature correspondences to the 3D models to estimate the pose and augment the video. By contrast, PARIS performs all processing locally [26, 27]. Wagner *et al* report a PDA AR solution [23] that perform local processing by leveraging the ARToolkit, whereas PARIS uses a VRML solution.

Ressler *et al* [19] describe a technique for integrating speech synthesis output within VRML, however the integration of speech recognition is not considered at all. Mynatt *et al* [14] describe, Audio Aura, to provide office workers with rich auditory cues (via wireless headphones) within the context of VRML for describing the current state of the physical objects that interest them. By contrast, PARIS supports speech in and out for dialog using VoiceXML. In addition, neither approach from Ressler nor Mynatt consider PDAs.

## 3. ARCHITECTURE

Important technology considerations were to embrace international standards where suitable, execute on a commercially available PDA, and also to leverage any appropriate products and schemes used in contemporary plants. As such, PARIS supports VRML and VoiceXML for the graphic and speech media. The PDA device used was a regular Compaq iPAQ Pocket PC equipped with 200MHz processor, 64Mb memory, a Compact Flash camera and an 802.11b wireless card.

The PARIS framework comprises five significant components or subsystems: Augmented reality management unit, VRML engine, VoiceXML Lite [11, 5], Tracking and localization and Plant management communication. A high-level functional view of PARIS can be seen in figure 2. The inputs are on the left, the outputs are on the right, and wireless communication with the plant management system is below. The following sections explain the function and interaction of these components.

The central command and control center of the PARIS architecture is the Augmented Reality (AR) Management Unit (figure 3) which is responsible for orchestrating all interaction between VoiceXML Lite, the VRML engine and the Tracking and Localization components. The VRML Manager is responsible for initializing and instructing the engine to load the appropriate VRML world. In addition, the VRML Manager is responsible for coordinating and

synchronizing any updates as the technician interacts with the scene. The Video Manager provides an interface to the tracking and Localization software. It initializes the camera, makes appropriate calls to and receives events from the Localization and Tracking component to perform the marker detection. The Voice Manager governs communication with the VoiceXML Lite subsystem. It drives VoiceXML Lite, determining the appropriate VoiceXML file to be loaded based upon the purpose of the user interaction. The Voice Manager gathers parameters from the user during the speech dialog and forwards messages to the VRML Manager for possible visual display.

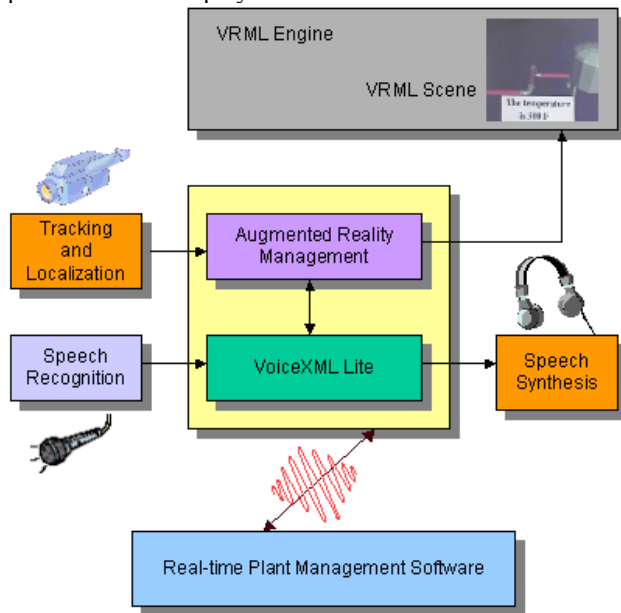


Figure 2: High-level view of PARIS.

The architecture is driven entirely by the inputs, either from the input sensors or from the user. The typical flow is described below. The tracking and location component processes the video searching for visual markers. Upon successful detection of a marker, the Video Manager receives an event indicating the unique identity of the marker. The tracking and localization algorithm is then stopped, as it degrades the performance of the speech interaction and visualization. The Video Manager then communicates this information to the Voice Manager and the VRML Manager. The Voice Manager instructs VoiceXML Lite to load the VoiceXML file associated with this marker and to begin the speech dialog. The system then engages the user in a speech dialog. As each spoken form input item is elicited from the technician, an embedded JavaScript function is executed to pass the data onto the Voice Manager. The Voice Manager creates a message into which the input data is placed. This data represents information pertaining to the nature of the task and the entities affected. This message is then written to a message

queue. This process is repeated until all of the required speech inputs have been gathered. Unless already loaded, the VRML Manager instructs the VRML engine to load the VRML world associated with this marker and transitions the virtual camera position to the corresponding viewpoint. The VRML Manager then periodically polls the message queue for pending messages from the Voice Manager. Just as HTML and VoiceXML can contain JavaScript, so too can VRML. In addition to geometrical information, the VRML world is imbued with an extensible collection of JavaScript nodes with parameterized functions for performing visual actions to change the VRML scene. The information extracted from a message and maps onto the parameters of these JavaScript functions. The VRML Manager interacts with the VRML engine to set each JavaScript parameter in the VRML node(s) and then invokes the appropriate JavaScript function(s) to perform various visual action(s). Among the visual actions currently provided are highlighting nodes, overlaying nodes, displaying signposts containing textual descriptions, etc.

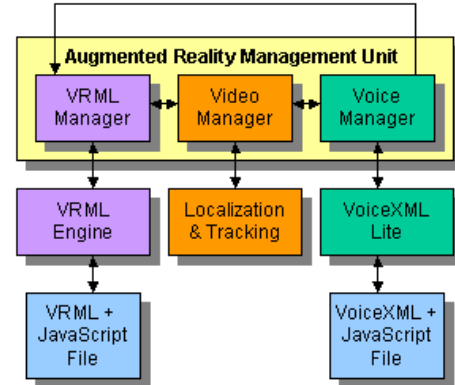


Figure 3: Augmented Reality Management Unit.

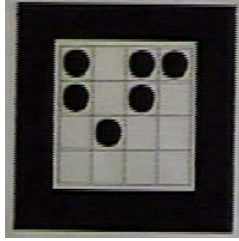
The Video Manager repeats the cycle by re-starting the localization and tracking component.

#### 4. VISION-BASED LOCALIZATION

Coded visual markers are used to support motion tracking and localization, and are used in many industrial sites for photogrammetry. In many cases, marker positions are measured to millimeter precision and stored in databases. Algorithms exist for computing the 3D position and orientation of a camera relative to markers. An example of the coded visual markers employed by PARIS can be seen in figure 4. The rectangular frame is used for the marker detection and for image correspondences. Using the 4x4 coding matrix, more than ten thousand uniquely coded visual markers can be generated. Each marker provides at least eight feature points for image correspondences.

The visual marker-based localization is implemented through motion tracking and camera calibration. A homography-based camera calibration algorithm exploits

the correspondence between a set of coplanar points and their images to estimate the position and orientation of the camera. With every coded visual marker pre-registered within the global coordinate system, the 3D position and orientation of the camera attached to the PDA can be thus determined [27]. Below is a brief description of the camera calibration algorithm.



**Figure 4: An example of a coded visual marker.**

The pinhole camera model describes the relationship between a 3D point,  $\mathbf{M} = [X, Y, Z, 1]^T$ , and its 2D projection,  $\mathbf{m} = [u, v, 1]^T$ , on the image plane as

$$s \mathbf{m} = \mathbf{A} [\mathbf{R} \mathbf{t}] \mathbf{M} \quad (1)$$

where  $s$  is a scaling factor,  $\mathbf{R} = [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{r}_3]$  the  $3 \times 3$  rotation matrix,  $\mathbf{t}$  the  $3 \times 1$  translation vector, and  $\mathbf{A}$  the

camera intrinsic matrix given by  $\mathbf{A} = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix}$ ,

with  $(u_0, v_0)$  being the coordinates of the camera optical center on the image plane,  $\alpha$  and  $\beta$  the focal lengths in image  $u$  and  $v$  directions, and  $\gamma$  the skewness of the two image axes. Since all 3D points are on the model plane, we construct the marker coordinate system with  $Z=0$ . Thus equation.(1) can be rewritten as

$$s \mathbf{m} = \mathbf{A} [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{r}_3 \ \mathbf{t}] [X \ Y \ 0 \ 1]^T$$

$$= \mathbf{A} [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{t}] [X \ Y \ 1]^T = \mathbf{H} [X \ Y \ 1]^T \quad (2)$$

or

$$s \mathbf{m} = \mathbf{H} [X \ Y \ 1]^T \quad (3)$$

where  $\mathbf{H}$  is the  $3 \times 3$  homography describing the projection from the marker plane to the image plane. We note

$$\mathbf{H} = [\mathbf{h}_1 \ \mathbf{h}_2 \ \mathbf{h}_3] = \lambda \mathbf{A} [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{t}] \quad (4)$$

Since at least 8 pairs of correspondences can be obtained from each marker, the homography  $\mathbf{H}$  can be determined up to a scaling factor. In many cases, the intrinsic matrix  $\mathbf{A}$  is given from off-line camera calibration, then the rotation matrix  $\mathbf{R}$  and translation vector  $\mathbf{t}$  can be obtained. The final results are then optimized by minimizing the following function for a set of  $n$  images, each with  $m$  known coplanar 3D points:

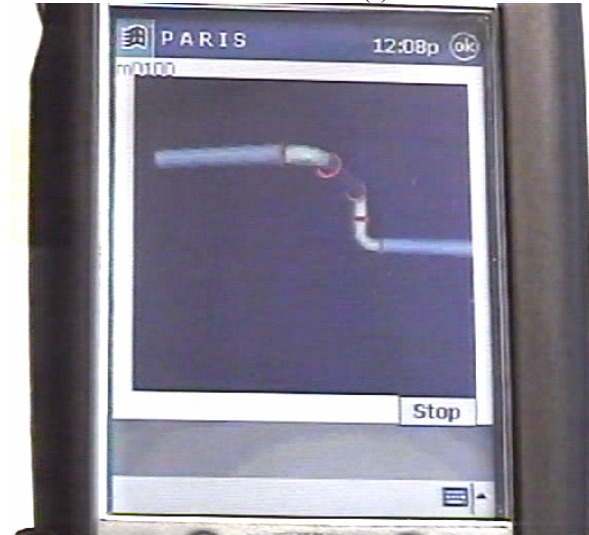
$$\sum_{i=1}^n \sum_{j=1}^m \| \mathbf{m}_{ij} - \mathbf{m}'(\mathbf{A}, \mathbf{R}_i, \mathbf{t}_i, \mathbf{M}_j) \|^2 \quad (5)$$

where  $\mathbf{m}'(\mathbf{A}, \mathbf{R}_i, \mathbf{t}_i, \mathbf{M}_j)$  is the projection of point  $\mathbf{M}_j$  in image  $i$ . This nonlinear optimization problem can be solved with the Levenberg-Marquardt algorithm [17].

Previous work [3, 7, 8, 9] suggests that a PDA is not powerful enough to perform the real-time image processing and augmentation. The implementation of PARIS reported in this paper proves the contrary, as the PDA is the only computer used for processing. The video analysis, marker-based tracking and localization, image augmentation and AR visualization all execute efficiently on the PDA. The current implementation provides detection of the markers in real-time: image acquisition and marker detection performs well at 10 frames per second or more, depending on other processes being managed by the PDA. It is also able to estimate the position and orientation of the PDA, and therefore its user, relative to the environment. The technician can stand approximately 10 feet away from the marker.

## 5. INTERACTION AND VISUALIZATION

The plant components designated to be visualized and speech-enabled are each labeled with unique visual markers, as can be seen in figure 1. Corresponding VRML worlds and VoiceXML scripts must be generated or created. The collection of JavaScript functions that provide the visual actions must be embedded or referenced in the VRML file(s). JavaScript for enqueueing a message must also be embedded or referenced in the VoiceXML file(s).



**Figure 5: 3D visualization of a pipe assembly with the joints highlighted in red.**

Current maintenance practice ranges from completing paper-based forms through to using various flavors of mobile computers for inputting values into fields of server-generated HTML forms. Instead of server-generated HTML forms, PARIS could retrieve and process server-generated VoiceXML forms. PARIS facilitates the quasi-synchronized rendering of the synthesized speech output with its visual counterpart. This feature is necessary for confirming interactions in noisy environments.

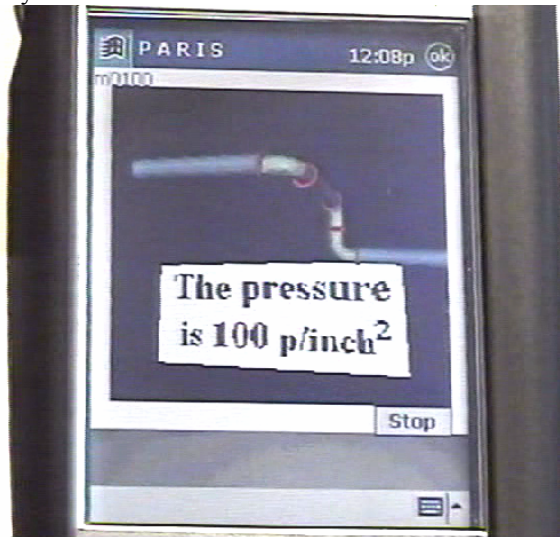


Figure 6: A text label is displayed as a result of a speech-driven query for the pressure in the joints.

A scenario was developed in order to test and evaluate the framework in the lab environment. A maintenance technician patrols the plant with her PDA and approaches the first piece of equipment scheduled for repair. PARIS automatically localizes her and displays the pipe assembly, as seen in figure 5.

Struggling to identify the joint at fault, she asks the system to highlight the joint. PARIS then highlights in red the joint and repositions the viewpoint for greater clarity. The technician performs the repair. Once the faulty joint has been replaced and the pipe assembly reconnected and enabled, the technician begins a series of tests to verify that the fault has indeed been corrected satisfactorily. She says “Pressure”, to which the framework issues a wireless HTTP query to the Siemens WinCC plant automation software. Upon receipt of the pressure information PARIS announces, “Current pressure is 100 pounds per square inch” while displaying a text label containing the same information (figure 6).

She walks on and climbs a nearby ladder to observe the connection of the pipe assembly to a container tank. PARIS again localizes her and transitions the viewpoint to reflect her new location. She then issues a final check by requesting “Temperature”, to which the framework again queries WinCC for the real-time value. Upon receipt PARIS announces, “Current temperature is 300

degrees Celsius” while displaying a text label confirming this value. This can be seen in figure 7.



Figure 7: The pipes are highlighted in red, and a label is displayed as a result of a speech-driven query for the temperature in the pipes.

## 6. FUTURE WORK AND CONCLUSIONS

While a number of tracking technologies have been proposed in the literature, Klinker *et al* [13] recognizes that the most successful indoor tracking solutions will comprise two or more tracking technologies to create a holistic sensing infrastructure able to exploit the strengths of each technology. We subscribe also to this philosophy, hence current work involves integrating supplementary localization technologies [10], providing support in areas where visual markers are either not present or cannot be detected effectively.

It possible to estimate from the video the position and orientation of the PDA, and therefore its user, relative to the environment. As the VRML worlds can become too large in size for the PDA, current work includes a scheme for downloading and caching of partial VRML worlds. By transmitting to a server the marker identity and the user’s location relative to the identified marker, the corresponding partial 3D worlds can be returned. Extensions to support mobile collaborative fault diagnosis are also in development. These include the ability to support a full duplex SIP/RTP voice-over-IP channel and a shared VRML browsing session with a remotely located expert.

Industrial service and maintenance is by necessity a mobile activity. The aim of this research is to improve the automated support for the technician in this endeavor. As such, a framework was developed called PARIS (PDA-based Augmented Reality Integrating Speech) that executes entirely on a commercially available PDA equipped with a small camera and

wireless support. Real-time computer vision-based techniques are employed for automatically localizing the technician within the plant. Once localized, PARIS offers the technician a seamless multi-modal user interface juxtaposing a VRML augmented reality view of the industrial equipment in the immediate vicinity and initiates a context-sensitive VoiceXML speech dialog concerning the equipment.

Important technology considerations were to embrace international standards where suitable (VRML and VoiceXML), execute on a commercially available PDA, and also to leverage any appropriate products and schemes used in contemporary plants. As such, PARIS supports VRML and VoiceXML for the graphic and speech media.

Although industrial mobile service and maintenance has provided the application context throughout this paper, the authors are exploring potential applicability of the technology in other vertical markets, such as healthcare, tourism and building information systems.

## 7. REFERENCES

- [1] BMBF funded AR-PDA Project, <http://www.ar-pda.de/>
- [2] Beigl, M., Memoclip: A Location-based Remembrance Appliance, *Journal of Personal Technologies*, 4(4):230-234, Springer Press, 2000.
- [3] Bertelsen, O., Nielsen, C., Augmented Reality as a Design Tool for Mobile Interfaces, *Proceedings of Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, ACM Press, pages 185-192, New York, 2000.
- [4] Dai, F., *Virtual Reality for Industrial Applications*, Springer-Verlag, 1998.
- [5] Eberman, B., Carter, J., Meyer, D. and Goddeau D., Building VoiceXML Browsers with OpenVXI, *Proceedings of the 11th ACM International World Wide Web Conference*, Hawaii, USA, pages 713-717, May, 2002.
- [6] Fitzmaurice, G., Zhai, Z. and Chignell, M., Virtual Reality for Palmtop Computers, *ACM Transactions on Office Information Systems*, 11(3):197-218, July, 1993.
- [7] Gausemeier, J., Freund, J., Matyszczok C., Bruederlin B., Beie D., Development of a Real Time Image-Based Object Recognition Method for Mobile AR-Devices, *Proceedings of International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction*, Africa, February 2003.
- [8] Geiger, C., Kleinjohann, B., Reimann, C. and Stichling, D., Mobile AR4All, *Proceeding of IEEE and ACM International Symposium on Augmented Reality*, New York, 2001.
- [9] Geiger, C., Paelke, V., Riemann, C., Rosenbach, W., Stoecklein, J., Testable Design Representation for Mobile Augmented Reality Authoring, *Proceeding of the IEEE and ACM International Symposium on Mixed and Augmented Reality*, Darmstadt, Germany, 2002.
- [10] Goose, S., Wanning, H. and Schneider, G., Mobile Reality: A PDA-Based Multimodal Framework Synchronizing a Hybrid Tracking Solution with 3D Graphics and Location-Sensitive Speech Interaction, *Proceedings of the ACM 4th International Conference on Ubiquitous Computing*, Göteborg, Sweden, pages 33-47, September, 2002.
- [11] Goose, S., Kodlahalli, S. and Lukas, K., VoiceXML Lite: A Standards-based Framework for Speech-Enabling Applications Executing on Commodity Networked Devices, *Proceedings of the IEEE International Conference Distributed Multimedia Systems*, Florida, USA, September, 2003.
- [12] Hinkley, K., Pierce, J., Sinclair, M. and Horvitz, E., *Sensing Techniques for Mobile Interaction*, ACM UIST, San Diego, USA, November 2000.
- [13] Klinker, G., Reicher, T. and Bruegge, B., Distributed User Tracking Concepts for Augmented Reality Applications, *Proceedings of ISAR 2000*, Munich, Germany, pages 37-44, October, 2000.
- [14] Mynatt, E., Back, M., Want, R., Baer, M. and Ellis, J., Designing Audio Aura, *ACM International Conference on Computer Human Interaction*, Los Angeles, USA, pages 566-573, 1998.
- [15] Newman, J., Ingram, D. and Hopper, A., Augmented Reality in a Wide Area Sentient Environment, *Proceedings of IEEE International Symposium on Augmented Reality (ISAR)*, pages 77-86, October 2001.
- [16] Nilsson, J., Sokoler, T., Binder, T. and Wetcke, N., Beyond the Control Room: Mobile Devices for Spatially Distributed Interaction on Industrial Process Plants, *Proceedings of the Second International Symposium on Handheld and Ubiquitous Computing*, Bristol, U.K., pages 30-45, September 2000.
- [17] Press, W., Teukolsky, S., Woo, M. and Flannery, B., *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1992.
- [18] Rekimoto, J. and Ayatsuka, Y., *Cybercode: Designing Augmented Reality Environments With Visual Tags, Designing Augmented Reality Environments*, 2000.
- [19] Ressler, S. and Wang, Q., Making VRML Accessible for People with Disabilities, *ASSETS 98*, Marina del Rey, USA, pages 50-55, April 1998.
- [20] Smalagic, A. and Bennington, B., Wireless and Mobile Computing in Training Maintenance and Diagnosis, *IEEE Vehicular Technology Conference*, Phoenix, AZ, May 1997.
- [21] Voice Browser Working Group, <http://www.w3.org/Voice>
- [22] VRML97 Specification, ISO/IEC 14772-1:1997
- [23] Wagner, D., and Schmalstieg, D., First Steps Towards Handheld Augmented Reality, *ISWC*, New York, USA, October 2003.
- [24] Want, R., Hopper, A., Falcao, V. and Gibbons, J., The Active Badge Location System, *ACM Transactions on Information Systems*, 10(1):91-102, 1992.
- [25] Zhai, S., Milgram, P. and Drasic, D., An Evaluation of four 6 Degree-of-Freedom Input Techniques, *ACM Conference on Human Factors in Computer Systems*, Amsterdam, Netherlands, 1993.
- [26] Zhang, X. and Navab, N., Tracking and Pose Estimation for Computer Assisted Localization in Industrial Environments, *IEEE Workshop on Applications of Computer Vision*, pages 214- 221, 2000.
- [27] Zhang, Z., Flexible Camera Calibration by Viewing a Plane From Unknown Orientations, *IEEE International Conference on Computer Vision*, pages 666-673, 1999.