VISUAL SCENE AUGMENTATION FOR ENHANCED HUMAN PERCEPTION

Daniel Hahn, Frederik Beutler and Uwe D. Hanebeck
Intelligent Sensor-Actuator-Systems Laboratory
Institute of Computer Science and Engineering
Universität Karlsruhe (TH)
Karlsruhe, Germany

Keywords: Augmented Reality, Human-Machine-Interface.

Abstract: In this paper we present an assistive system for hearing-impaired people that consists of a wearable microphone array and an Augmented Reality (AR) system. This system helps the user in communication situations, where many speakers or sources of background noise are present. In order to restore the "cocktail party" effect multiple microphones are used to estimate the position of individual sound sources. In order to allow the user to interact in complex situations with many speakers, an algorithm for estimating the user's *attention* is developed. This algorithm determines the sound sources, which are in the user's *focus of attention*. It allows the system to discard irrelevant information and enables the user to focus on certain aspects of the surroundings. Based on the user's hearing impairment, the perception of the speaker in the *focus of attention* can be enhanced, e.g. by amplification or using a speech-to-text conversion.
A prototype has been built for evaluating this approach. Currently the prototype is able to locate sound beacons in three-dimensional space, to perform a simple focus estimation, and to present floating captions in the Augmented Reality. The prototype uses an intentionally simple user interface, in order to minimize distractions.

1 INTRODUCTION

Hearing impairments have grave consequences on a person's social life. Everyday tasks and social interaction depend on spoken language, a form of communication from which the hearing impaired are often cut off. Speechreading skills and assistive technology (such as conventional hearing aids or cochlear implants) can remedy the problems to a degree. These approaches, however, can only assist a person's remaining cognitive capabilities. They do not aim at restoring the complex functions of human hearing, and thus do not work well in complex situations. These limitations of conventional hearing aids are best illustrated by their failure restoring the "cocktail party" effect – the phenomenon that allows us to focus on a certain speaker and put other speakers and noises to the background.

For this reason, hearing-impaired people face an increased danger of social isolation. This is especially true for those who became deaf later in life, and have grown into a world of spoken language. While these people are able to express themselves orally, they are often unable to understand what is said. Therefore they may avoid social interactions, or at least situations with which their hearing aids cannot cope.

The system in this paper was born out of an idea to overcome these limitation: A novel hearing aid which would combine a wearable microphone array with an augmented reality headset. By using multiple redundant microphones, it will be possible to estimate the position of the individual sound sources surrounding the user, and to isolate their signals from the background noise. The sound information will then be visually presented to the user in the augmented reality headset, using the mode of presentation that is best suited for the user's abilities.

In this paper we will present a first prototype of the system, as well as methods for selecting the audio information and integrating them into the user's reality.

Augmented reality systems integrate virtual objects into the user's real surroundings (Azuma, 1997). In the case of the hearing aid, the sound sources are augmented with virtual representations of their content. In the case of spoken language, these will most likely be textual representations of the speech, which are attached to the respective speaker (e.g. speech bubbles).

Augmented reality can be regarded as a new form

Hahn D., Beutler F. and D. Hanebeck U. (2005).
VISUAL SCENE AUGMENTATION FOR ENHANCED HUMAN PERCEPTION.
In *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics - Robotics and Automation*, pages 146-153
DOI: 10.5220/0001173701460153

Figure 1: Mockup of a possible user interface.

of user interface and requires new forms of user interaction. There have been attempts to bring elements of traditional GUI interfaces into the AR, like the Studierstube project (Schmalstieg et al., 2002). Other researchers tried to use novel approaches, incorporating tangible objects into the interface (Tan et al., 2001). All these attempts regard the AR system as a tool that requires constant interaction. There have been some attempts on information-only systems, like the emergency room prototype by Kaufman, Billinghurst et. al. (Kaufman et al., 1997). However, no accepted design rules for such systems seem to have evolved.

A problem for the development of the new hearing aid is the correct *registration* of the augmented world. The virtual elements have to be perfectly aligned with the virtual word in order to be convincing. This is a problem that has not been completely solved yet (Azuma, 1997). For the hearing aid the registration does not need to be as perfect as in other applications, but it must work without the help of external tracking devices or visual markers, as used in the ARToolkit (Billinghurst et al., 2004).

In order to work in complex situations, and to restore the cocktail party effect, the system needs to know which speakers need to be represented in the AR, and which sounds need to be suppressed. Critical to this is the concept of human *attention*. The mechanisms of attention, which, in people with normal hearing, functions without conscious effort, have to be restored.

This requires that the system contains a model of the user's attention. Such a model will allow the system to intelligently decide which information should be augmented into the user's reality, and which should be discarded. This will enable the system to reduce the amount of information to a level which the user can easily understand.

The information needs to be presented through an intuitive and non-obtrusive interface. The interface itself must not inhibit the user in any way or interfere with his everyday tasks. This precludes the use of complex and graphics-heavy interfaces. A major part of the interface design will be the visual representation of spoken language. While different approaches are possible, a textual representation seems to be the most intuitive for the first prototype (Figure 1).

Some ideas for such an interface can be found in a class of multimodal user interfaces, known as "attentive" or "perceptual", which pioneered the use of human attention in user interfaces. Examples are Vertegaal's *Attentive Interfaces* (Vertegaal, 2002a) or Pentland's *Perceptual Intelligence* (Pentland, 2000).

These systems attempt to monitor the attention of their users, in order to interact with them more intelligently. An example is Vertegaal's attentive cell phone, which observes the user's conversational partner to determine whether a call should be put through (Vertegaal et al., 2002).

The primary method of getting information about the user's attention is through the observation of the gaze (Sibert and Jacob, 2000), (Vertegaal, 2002b). Since professional gaze-tracking equipment is bulky and expensive, many researchers attempt to build simple eye-tracking tools, using off-the-shelf hardware like webcams. An example is Vertegaal's "eyes" system (Shell et al., 2003). Stiefelhagen tries to ascertain the gaze by tracking the head pose, with surprisingly good results (Stiefelhagen, 2002).

Attentive interfaces are usually seen as extensions of graphical user interfaces (GUIs) (Vertegaal, 2003). They are supposed to mediate the human-computer-interaction (Shell et al., 2003). In that capacity their role can be describe as that of an "intelligent observer" with a social awareness of communication.

These systems do not necessarily attempt to fully model the user's attention, as it is necessary for the hearing aid.

The structure of the paper is as follows. In Section 2 models for human attention are described. In Section 3 a model for the human attention is deduced. In Section 3.1 we describe how the attentional state is estimated. In Section 3.2 the target selection and processing are described. The prototype of the system is presented in Section 4. In Section 4.1 an overview over the system architecture is given. Further in Section 4.2 the hardware setup and in Section 4.3 the software setup are described. Experiments with the prototype are shown in Section 4.4. In Section 5 the results of the experiments are presented. Conclusions and some details on future investigations are given in Section 6.

2 HUMAN ATTENTION

Attention is the ability to selectively focus on certain parts of one's surroundings, while disregarding the other parts. Attention has often been compared to a spotlight, which selectively illuminates objects in a dark room.

Human attention has been extensively studied by cognitive psychologists, and there's a wealth of literature available on the issue (Chun and Wolfe, 2001). There are two prevailing schools of thought within the literature: *Filter* or *attenuation* theories, as propagated by Broadbent (Broadbent, 1958) and Treisman (Treisman and Gelade, 1980), assume that attention works like a filter. Unneeded perceptions are either removed or toned down, and do not enter consciousness.

Resource models, on the other hand, propose that attention is created by the distribution of limited attentional resources (Cohen, 2003). Those processing resources can be allocated to different perceptions, which allows them to be consciously perceived. Those perceptions for which no resources are available will be discarded.

Both models can be used to explain the results of psychological experiments (Cohen, 2003). We will primarily use the resource model, since it makes it easy to describe attention in computational terms.

Cognitive psychology has revealed many more mechanisms of attention (Chun and Wolfe, 2001):

- The spotlight of attention can be divided, multiple objects can be attract attention at the same time. However, the overall performance always remains the same.
- Attention can be shifted through a conscious effort. However, it can also be drawn by certain features of the environment. For example, a blinking light will immediately draw a person's attention. This kind of attention shift occurs automatically and requires no conscious effort. This property of attention is exploited in image processing algorithms which attempt to imitate the visual attention, for an example see (Backer and Mertsching, 2003).
- While attention has spatial properties, it can also work on whole objects. This indicates that objects can be identified in a preattentive processing stage.

3 A MODEL OF ATTENTION

The attention model developed for the hearing aid assumes that the user's attention can be directed at a number of possible *targets*. Each of these targets is a distinct entity corresponding to an object in the real world. For example, a speaker in a room would be a possible target for the user's attention.

Each target is attributed with a *target description*. The descriptions contains the raw sensor data from that target, and may also contain semantic information that can be used for estimating the user's attentional focus. A target description for a speaker may consist of the raw audio data from this speaker and the speaker's position relative to the user.

The *attentional state* of the user is the distribution of the user's attention over the existing targets. The distribution is expressed, for each target, as the probability that the target is the user's primary focus of attention. This model is consistent with the psychological results which indicate that attention is directed at objects, rather than abstract features.

For estimating the attentional state the possible targets have to been detected in the sensor information, and each target's sensor data is extracted separately. This may seem like an excessive burden on the preprocessing stage. However, in the case of the hearing aid, advanced audio processing has to be an integral part of the system anyway. All sound sources will have to be identified and localized, and there has to be a possibility to enhance each sound source separately or feed it to a speech recognition system.

3.1 Estimating the attentional state

The model for the user attention consists of a number of rules. By assigning a probability to each target the algorithm creates an estimate of the user's attentional state. Since the rules are interchangeable, different approaches may be evaluated. This is necessary since psychological experiments suggest a wealth of approaches, but it is often unclear how they will behave in real-life systems.

There are two basic approaches to determine the user's attentional state. One is by predicting the attention based on the user's current perceptions. The other is to monitor the user's behavior in order to find out where the attention is directed.

Figure 2 shows a coarse overview over the mechanisms of the algorithm. We assume that the user receives perceptions or *stimuli* from the world and, depending on his current attentional state, reacts to those stimuli. The stimuli are recorded by sensors and transformed into target descriptions in a preprocessing stage. Based on the target descriptions and the user model the user's most likely attentional state is estimated.

Simultaneously, the user's reactions are monitored. Through the reactions, the system may observe the user's attentional state. Any differences between the estimated and observed state are fed back into the model.

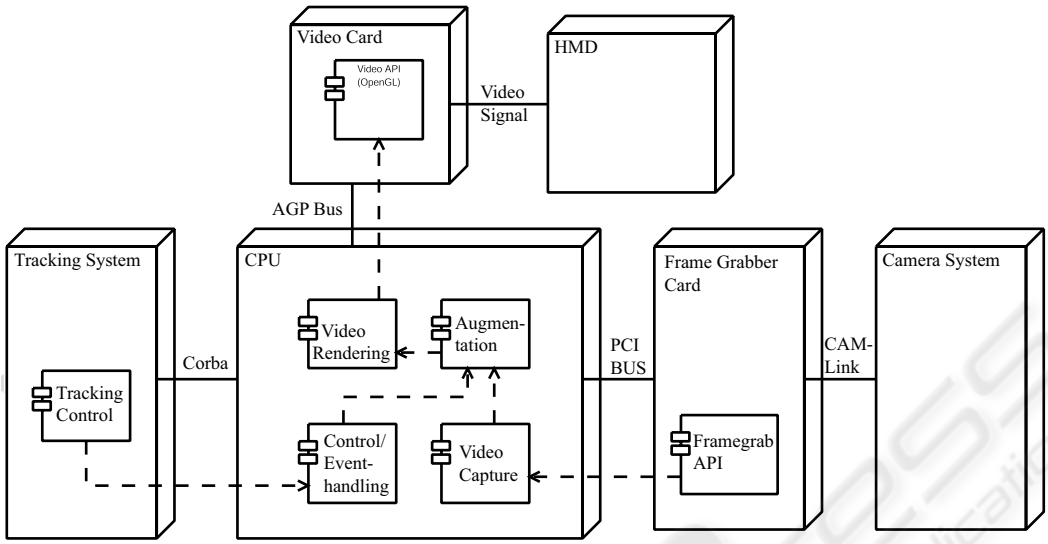

Figure 3: Components of the Ve system.

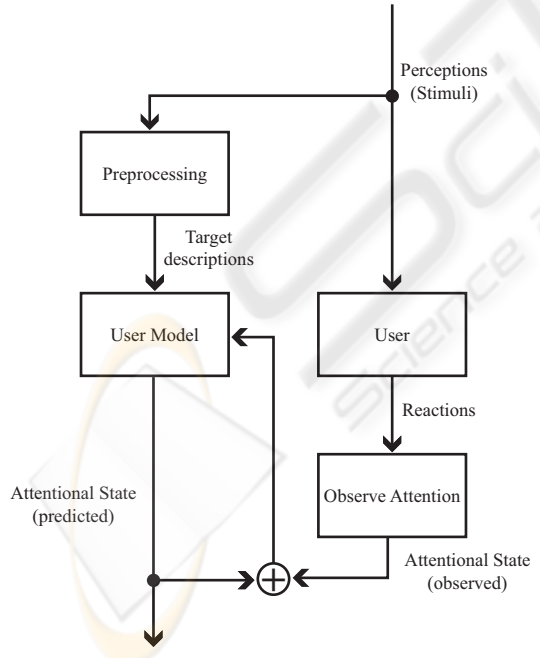

Figure 2: Algorithm for estimating the attentional state.

Observing attention If a person shifts his or her attention, this shift will often result in a behavior that can be registered by the system. This approach will be especially useful monitoring the users conscious, *extrinsic* attention – since the extrinsic attention is guided by the user's will, it cannot be expressed through fixed rules.

The best known method observing the user's attention is by tracking the user's gaze. Since the classical experiments of Yarbus it is know that there is a close connection between the eye movements and a person's focus of attention (Yarbus, 1967). This connection has been exploited many times, especially for user interface designs. Vertegaal calls eye-based interaction an '*almost magical window into the mind of the user*' (Vertegaal, 2002b).

For the hearing aid we simply use the *head posture*, which can be easily obtained. Stiefelhagen has shown that head posture and gaze are well correlated, and has used this fact with great success in a conferencing system (Stiefelhagen, 2002).

Apart from gaze tracking, there are very few methods that can be used to observe the user's attention. While gestures and body posture may be indicators of attention, they are much more difficult to evaluate and not nearly as precise as gaze tracking.

Predicting attention Based on the forward user model, which consists of a set of rules, the user's attentional state given the current perceptions is predicted. Any assumption about how a certain perception changes the user's attention may be used as a rule

in the model.

The model functions like the user's unconscious or *intrinsic* attention; it gives an estimate of the user's attentional state that is based on the probability of each object to draw the attention. While providing a structured framework, it is open in the sense that different approaches and rules may be evaluated within the same framework.

An example for attention-predicting systems are the image processing techniques that attempt to locate a probable focus of attention within an image (Backer and Mertsching, 2003), (Draper and Lionelle, 2003). Such an approach may be part of our user model, but the user model is not limited to a single approach.

For example, the system may decide that a person speaking is much more likely to be the focus of attention than a person not speaking (Stiefelhagen, 2002) and adjust the score accordingly. Other indicators that may be used are the object's distance from the user or the volume of an audio source. The system may even scan the audio streams for trigger words, or make assumptions about the current social situation, in order to better estimate the attentional state.

The current prototype does not yet contain a sophisticated user model, but relies on the user's head posture to observe the current attentional state. A more complex user model will be included in future versions, and we assume that the complexity of the model will be dictated by the demands of the system and the processing power available.

3.2 Target selection and processing

Once the attentional state is known, the system may select a number of targets for presentation. The number of targets chosen will depend on the actual situation; there may be circumstances where it is necessary to have more than a single focus of attention. However, if too many targets are augmented, the user easily becomes confused.

In order to present a target, the target description including the raw sensor data will be transformed into a format that is intelligible to the user. The format depends on the user. For users which are hard of hearing the separated audio signal are amplified and emitted through earphones. For a deaf user, on the other hand, a speech-to-text conversion through a speech recognition system can be used. Since the target description will already contain the separated audio signal from a single source, reliable speech recognition should be feasible.

Other transformations are also possible: Audible signals (e.g. a ringing phone) could be transformed to pictograms or less important speakers could be represented through symbols. It is even imaginable that the system transforms the speech to a sign language representation.

4 VE PROTOTYPE

A prototype of the hearing aid was built to evaluate the claims made in this paper, and to improve the methods for estimating the attention. At the core of the prototype is a custom-built AR system, which was developed at the Intelligent Sensor-Actuator-Systems laboratory.

4.1 System architecture

The AR system is built upon the C++ class library *Ve*[1]. The library offers generic methods for video access, stereoscopic camera calibration and methods to augment the video streams.

A general overview of the system's architecture is shown in figure 3. At the core of the system is the Ve software. It captures the video feeds from the cameras and creates the virtual objects from the information received by the tracking subsystem. The video feeds and virtual objects are combined and presented to the user in a head mounted display (HMD).

4.2 Hardware setup

For the prototype, we built a *video see-through* AR system, using a commercially available HMD unit (Figure 4). Video see-through units capture the real world through cameras; the camera images are then fed into an opaque HMD. Compared to optical systems, where the user directly sees the surroundings through semi-transparent glasses, video systems offer a higher degree of flexibility: Every aspect of the display can be customized as needed.

The prototype's stereoscopic camera head is equipped with two high-resolution cameras by Silicon Imaging. They are connected to the controlling PC (Pentium 4, 2.8Ghz, Windows XP) by standard CAMLink frame grabber cards. The PC also contains an nVidia Quattro graphics card. The adapter has been chosen for it's ability to provide two separated digital video feeds to the HMD and because of the availability of high-quality optimised OpenGL drivers.

A high video framerate is necessary in order to create a realistic experience for the user. The system will currently provide a feed with 30 frames per second, with a resolution of 640×480 pixels, for each eye. While some performance gains may be archived through software optimizations, the limiting factor is currently the maximum throughput of the PCI bus. For the prototype, the full resolution of 1024×786

[1]Ve was the name of a norse god, who gave humanity speech and their external senses

Figure 4: Experimental AR System with audio tracking.

pixels may be achieved by using more advance technology; specialized hardware is most likely necessary for a final version of the hearing aid.

The HMD is a commercially available high-resolution system. The frame has been customized to provide mounting points for the camera head and the microphones of the tracking system. The headset is connected to a ceiling-mounted control unit. The setup allows the user to freely move about 3 meters in each direction. Through the headset, the users sees an augmented version of the surroundings, created by the Ve software.

Four microphones have been attached to the headset as part of a basic audio tracking system; they are connected to DSP board which is mounted to together with the HMD's control box. The tracking systems emulates some of the functionality of the planned microphone array. It is able to locate loudspeakers, which emitting a known signal; the position of the speakers is used for augmentation.

4.3 Software setup

The Ve software comprises three major components. The video capture subsystem, an event handling and control mechanism that connects to the tracking system, and a video rendering subsystem, which creates the augmented reality from the video stream and the tracking data.

The video capture subsystem provides a generic interface for accessing video sources. A Ve video module captures the video, using the hardware-specific APIs and protocols. The video stream is also decoded if necessary, and the API then provides pointers to the individual pictures in memory. The video module also sets up the hardware and provides an API to control the capture hardware (e.g. to set a different exposure on the camera). Currently only a video module for EPIX-based frame grabber cards exists, but further modules should be easy to implement. Each capture module runs in a separate thread, concurrently to other tasks.

The video output is rendered using the OpenGL video API. Camera pictures are displayed as textures, which works well with an appropriately optimized driver. Ve also requires the OpenGL implementation to support the "Imaging Subset", which contains several image manipulation functions.

The OpenGL API was also chosen for it's cross-platform availability. Ve has been compiled on Linux and MS Windows systems. The software is currently only used on Windows, however, since the Linux drivers are not sufficiently optimized.

Ve's video output is rendered in multiple layers to allow different modules to add to the augmentation. Each layer may contain it's own state information, and the Ve library offers some utility functions, such as stereoscopic calculations, for the layer modules.

Special modules are provided for camera calibration and AR registration. The calibration module provides an interface to the camera calibration methods of the OpenCV toolkit. The module will also compute the stereoscopic parameters of the camera setup from the OpenCV data. The registration module is determines a transformation between the coordinate systems of the real and the virtual world, by solving the underlying linear equation system. Some compensation of the non-linear distortions is also possible. Both registration and calibration can be done interactively from a simple HMD interface.

Ve contains a simple event handling mechanism. Ve modules may subscribe to events created by other modules, and thus react to changes in the environment. Currently the event handling is used for control information and position updates.

The tracking system is not part of Ve. Tracking information is provided to Ve through CORBA function calls. The Ve part of the tracking subsystem is only a stub which transforms the position updates into Ve events and notifies AR modules of the update. An augmentation module may then react to the update and add an virtual object to the user's view. For testing purposes the tracking mechanism of the ARToolkit can also be used as a tracking module.

4.4 Experimental setup

The prototype is set up for an interactive simulation of the intended AR interface. The AR system is connected to the audio tracking system, which monitors the position of two loudspeakers in the room. Each of

those loudspeakers represents a human speaker; predefined "speech messages" are placed near them in the augmented reality.

The "speech messages" are rendered according to the attention model presented earlier in this paper: Speakers near the center of the screen are augmented by large text messages, while speakers at the periphery of the visual field only get small messages or are not augmented at all. The speaker that is currently within the focus of attention can also be augmented by a translucent "focus marker" or crosshairs.

In the prototype, the coordinates of the speakers are manually registered to the virtual world. The stereoscopic parameters from the camera calibration are used to create the stereoscopic images of the text object.

The prototype was evaluated by about 10 persons, both male and female and at the age of 20 to 40 years. The subjects were to explore the environment by turning their heads and moving about, while the loudspeakers could be moved by the experimenter. The subjects typically used the system for about 10 to 20 minutes.

5 RESULTS

Virtually all subjects were satisfied with the impression of the virtual world. The video see-through was described as sufficiently realistic, even though the resolution had been reduced to 640×480 pixels. The small difference between the natural eye position and the camera position was not noticed after a short acclimatization period.

The subjects described the stereoscopic representation of the virtual objects as good, with the focus markers "hovering" in front of the loudspeakers. Due to the three dimensional view, the focus could be clearly marked in all three dimensions.

Use of the AR system was intuitive, the subjects were quickly able to select activate targets at will. The main drawback of the system during these initial tests was the weight of the head assembly, a problem which can be fixed by using more specialized hardware in future prototypes.

6 CONCLUSIONS

In this paper we presented the concept and prototype of a novel kind of hearing aid. The concept is based on the vision of a system incorporating advanced sensor technology, an augmented reality interface, and intelligent signal processing.

The system is aware of the user's attention, which allows it to customize the interface to the user's needs. The attention-driven interface also allows the system to address problems present in contemporary hearing aids, such as the reestablishment of the cocktail-party-effect.

We introduced a model of human attention, based on the findings of cognitive psychology. The model attempts to emulate attention and provides mechanisms to predict the user's behavior, as well as the possibility to correct the predictions by monitoring the user's behavior. The model aims at replacing lost attentive capacities, rather than at observing the user's attention externally. The system model is highly modular and can be extended for future prototypes.

A prototype has been built to evaluate the concepts. To this end, a custom AR system has been implemented with the modular class library *Ve*. It allows for quick changes of the user interface and the evaluation of multiple approaches.

First test runs showed the viability of the approach. Navigation within the augmented reality appeared intuitive and the users were easily able to direct their attention.

Future prototypes will include a refined attention-prediction, and it is assumed that the system will evolve into a small, wearable, and easy to use system.

REFERENCES

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

Backer, G. and Mertsching, B. (2003). Two Selection Stages Provide Efficient Object-Based Attentional Control for Dynamic Vision. In *International Workshop on Attention and Performance in Computer Vision*, pages 9 – 16, Graz, Austria.

Billinghurst, M. et al. (2004). ARToolkit Augmented Reality Toolkit. http://www.hitl.washington.edu/research /shared_space/download/.

Broadbent, D. (1958). *Perception and Communication*. Pergamon Press, London.

Chun, M. M. and Wolfe, J. M. (2001). Visual Attention. In *Blackwell's Handbook of Perception*, chapter 9, pages 272–310. Blackwell.

Cohen, A. (2003). Selective Attention. In *Encyclopedia of Cognitive Science*. Nature Publishing Group (Macmillan).

Draper, B. A. and Lionelle, A. (2003). Evaluation of Selective Attention under Similarity Transforms. In *International Workshop on Attention and Performance in Computer Vision*, pages 31–38, Graz, Austria.

Kaufman, N., Poupyrev, I., Miller, E., Billinghurst, M., Oppenheimer, P., and Weghorst, S. (1997). New Interface Metaphors for Complex Information Space Visualization: an ECG Monitor Object Prototype. In

Proceedings of Medicine Meets Virtual Reality, pages 131–140.

Pentland, A. (2000). Perceptual User Interfaces: Perceptual Intelligence. *Communications of the ACM*, 43(3):35–44.

Schmalstieg, D., Fuhrmann, A., Hesina, G., Szalavari, Z., Encarnacao, L. M., Gervautz, M., and Purgathofer, W. (2002). The Studierstube Augmented Reality Project. Technical Report TR-188-2-2002-05, Interactive Media Systems Group, Institute for Software Technology and Interactive Systems, Vienna University of Technology.

Shell, J. S., Selker, T., and Vertegaal, R. (2003). Interacting with Groups of Computers. *Communications of the ACM*, 46(3):40–46.

Sibert, L. E. and Jacob, R. J. K. (2000). Evaluation of Eye Gaze Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–288, The Hague, The Netherlands. ACM Press.

Stiefelhagen, R. (2002). Tracking Focus of Attention in Meetings. In *International Conference on Multimodal Interfaces*, page 273, Washington, DC, USA.

Tan, D. S., Poupyrev, I., Billinghurst, M., Kato, H., Regenbrecht, H., and Tetsutani, N. (2001). On-demand, In-place Help for Augmented Reality Environments. In *Ubicomp 2001*, Atlanta, GA, USA.

Treisman, A. and Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology*, (12):97–137.

Vertegaal, R. (2002a). Designing Attentive Interfaces. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, pages 23–30, New Orleans, La, USA. ACM Press.

Vertegaal, R. (2002b). What do the eyes behold for human-computer interaction? In *Proceedings of the symposium on Eye tracking research & applications*, pages 59–60. ACM Press.

Vertegaal, R. (2003). Introduction. *Commun. ACM*, 46(3):30–33.

Vertegaal, R., Dickie, C., Sohn, C., and Flickner, M. (2002). Designing Attentive Cell Phone using Wearable Eye-contact Sensors. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, pages 646–647. ACM Press.

Yarbus, A. L. (1967). Eye Movements During Perception of Complex Objects. In *Eye Movements and Vision*, pages 171–196. Plenum Press, New York, NY, USA.