Virtual Reality manuscript No. This is a pre-print of a paper accepted for publication in Virtual Reality. The final publication is available at www.springerlink.com. (DOI: 10.1007/s10055-010-0183-5)

An Augmented Reality Interface to Contextual Information

Antti Ajanki · Mark Billinghurst · Hannes Gamper · Toni Järvenpää · Melih Kandemir · Samuel Kaski · Markus Koskela · Mikko Kurimo · Jorma Laaksonen · Kai Puolamäki · Teemu Ruokolainen · Timo Tossavainen

Received: date / Accepted: date

Abstract In this paper we report on a prototype Augmented Reality (AR) platform for accessing abstract information in real-world pervasive computing environments. Using this platform, objects, people, and the environment serve as contextual channels to more information. The user's interest with respect to the environment is inferred from eye movement patterns, speech and other implicit feedback signals, and this data is used for information filtering. The results of proactive context-sensitive information retrieval are augmented onto the view of a handheld or head-mounted display or uttered as synthetic speech. The augmented information becomes part of the user's context, and if the user shows interest in the AR content the system detects this and provides progressively more information. In this paper we describe the first use of the platform to develop a pilot application, Virtual Laboratory Guide, and early evaluation results of this application.

Keywords augmented reality · gaze tracking · information retrieval · machine learning · pattern recognition

Authors ordered alphabetically

A. Ajanki · M. Kandemir · M. Koskela · M. Kurimo · J. Laaksonen · T. Ruokolainen
Aalto University, Department of Information and Computer Science, Finland
E-mail: firstname.lastname@tkk.fi

S. Kaski
Aalto University and University of Helsinki, Helsinki Institute for Information Technology HIIT, Finland
E-mail: samuel.kaski@hiit.fi

H. Gamper · K. Puolamäki · T. Tossavainen
Aalto University, Department of Media Technology, Finland
E-mail: firstname.lastname@tkk.fi

M. Billinghurst
The Human Interface Technology Laboratory New Zealand (HIT Lab NZ), University of Canterbury, Christchurch, New Zealand
E-mail: mark.billinghurst@canterbury.ac.nz

T. Järvenpää
Nokia Research Center, Tampere, Finland
E-mail: toni.jarvenpaa@nokia.com

1 Introduction

In pervasive computing systems, there is often a need to provide users with a way to access and search through ubiquitous information associated with real world objects and locations. Technology such as Augmented Reality (AR) allows virtual information to be overlaid on the users' environment (Azuma, 1997), and can be used as a way to view contextual information. However, there are interesting research questions that need to be addressed in terms of how to know when to present information to the user, and how to allow the user to interact with it. As Hendricksen et al (2002) point out, pervasive computing applications need to place few demands on the user's attention and be sensitive to context.

We are interested in the problem of how to efficiently retrieve and present contextual information in real-world environments where (i) it is hard to formulate explicit search queries and (ii) the temporal and spatial context provides potentially useful search cues. In other words, the user may not have an explicit query in mind or may not even be searching, and the information relevant to him or her is likely to be related to objects in the surrounding environment or other current context cues.

The scenario is that the user wears data glasses and sensors measuring his or her actions, including gaze patterns and further, the visual focus of attention. Using the implicit measurements about the user's interactions

with the environment, we can infer which of the potential search cues (objects, people) are relevant for the user at the current point of time, and augment retrieved information in the user's view (Figure 1).

This new augmented information forms part of the user's visual context, and once the user's interaction with the new context is measured, more fine-grained inferences about relevance can be made, and the search refined. Retrieval with such a system could be described as retrieval by zooming through augmented reality, analogously to text entry by zooming through predicted alternative textual continuations (Dasher system, Ward and MacKay (2002)) or image retrieval by zooming deeper into an image collection (Gazir system, Kozma et al (2009)).

To realize this scenario, several elements are needed. First, objects and people should be recognized as potential cues with pattern recognition methods. The relevance of these cues needs to be inferred from gaze patterns and other implicit feedback using machine learning methods. Second, context-sensitive information retrieval needs to operate proactively given the relevant cues. Finally, the retrieved information needs to be overlaid on the user's view with AR techniques and modern display devices. All this should operate such that the users are distracted as little as possible from their real work tasks.

In this paper, we present a hardware and software platform we have developed which meets these needs, and a demonstration prototype application created using the platform. This application is a *Virtual Laboratory Guide*, which will help a visitor to a university department find out about (i) teachers and teaching and (ii) researchers and research projects. The Virtual Laboratory Guide presents context-sensitive virtual information about the persons, offices, and artifacts that appear as needed and disappear when not attended to. As far as we know, the implementation of gaze-based relevance estimation for contextual information filtering, the use of augmented audio reality and their evaluations are the main novel contributions of this paper.

In the remainder of this paper, we first review earlier related work, and describe the lessons learned from this which our research builds on. Then we discuss the hardware and software platform we developed, and finally present the Virtual Laboratory Guide application and a user evaluation of the technology.

2 Related Work

In developing the wearable pervasive information retrieval system described above, our work builds on earlier research on augmented reality and contextual information retrieval.

2.1 Mobile Augmented Reality

Augmented Reality (AR) involves the seamless overlay of virtual imagery on the real world (Azuma, 1997). In recent years, wearable computers (Feiner et al, 1997) and even mobile phones (Henrysson and Ollila, 2004) have been used to provide a mobile AR experience. Using these platforms, researchers have explored how AR interfaces can be used to provide an intuitive user experience for pervasive computing applications. For example, Rauhala et al (2006) have developed a mobile phone based AR interface which communicates with ubiquitous sensors to show the temperature distribution of building walls.

Previous researchers have used wearable and mobile Augmented Reality systems to display contextual cues about the surrounding environment. For example, the Touring Machine (Feiner et al, 1997) added virtual tags to real university buildings showing which departments were in the buildings. A layer-based integration model for encompassing diverse types of media and functionality in a unified mobile AR browsing system is proposed in Lee et al (2009). A similar effect is created using the commercially available Layar[1] or Wikitude[2] applications for mobile phones, both of which provide virtual information tags on the real world.

These interfaces highlight the need to filter information according to the user's interest, and present it in an uncluttered way so that it is easy to interact with. In most cases, mobile AR interfaces require explicit user input specifying the topics of interest to the user. In our research we want to develop a system that uses unobtrusive implicit input from the user to present relevant contextual information.

2.2 Mobile Face Recognition

Gaze and face recognition provide important implicit cues about where the user is looking and whom he is interested in. Starner et al (1997) and Pentland (1998) describe on a conceptual level how wearable computers can be used as an ideal platform for mobile augmented reality, and how they can enable many applications, including face recognition. Pentland (2000) similarly points out how recognizing people is a key goal in computer vision research and provides an overview of previous work in person identification.

[1] http://layar.com/

[2] http://www.wikitude.org/

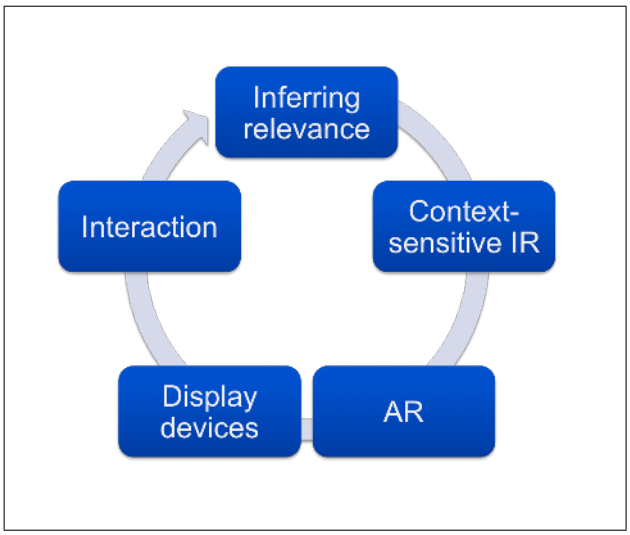

Fig. 1 Information retrieval (IR) is done in a loop where relevance of already augmented information (Augmented Reality, AR) and of the objects in the scene is inferred from user's observed interaction with them, then more information is retrieved given any contextual cues and the inferred relevance, and new information is augmented.

Several groups have produced prototype wearable face recognition applications for identifying people (Starner et al, 1997; Farringdon and Oni, 2000; Brzezowski et al, 1996; Iordanoglou et al, 2000). For example, Singletary and Starner (2001) have demonstrated a wearable face recognition system that uses social engagement cues to trigger the recognition process. They were able to achieve more than 90% recognition accuracy on tests with 300 sample faces. These systems have shown that it is possible to perform face recognition on a wearable platform; however, there has been little research on the use of AR and face recognition to trigger and present contextual cues in a pervasive computing setting. Our research uses face recognition to trigger contextual cues for finding context-sensitive information associated with the faces seen; the information is then shown using an AR interface.

2.3 Feedback in Contextual Information Retrieval

Information retrieval (IR) is an established field of study in which the fundamental question is how to find documents from a large corpus that best match the given query. The most visible everyday application of IR is Internet search. The query is typically a list of textual keywords or a collection of example items.

Object location search engines such as MAX (Yap et al, 2005) or Snoogle (Wang et al, 2008) let the user search for physical locations of objects that have been tagged with RFID transponders. However, they require that the user explicitly enters the object ID or terms describing the object as a query to the search engine.

Formulating a good search query can be difficult because of the cognitive effort required (Turpin and Scholer, 2006) or limitations of mobile input devices. A common way of decreasing the dependency on search queries is to utilize feedback from the user to guide the search. Feedback can be explicit, like in search engines which ask the user to rate documents in an initial result set, or it can be implicit feedback that is collected by observing the context and behavior of user.

In general, the use of context instead of or in addition to query-based IR is called contextual information retrieval. A simple example, again, is an Internet search engine that customizes the search result based on the location of the user (e.g., show only nearby restaurants). Contextual information retrieval takes into account the task the user is currently involved in, such as shopping or medical diagnosis, or the context expressed within a given domain, such as locations of the restaurants — see (Crestani and Ruthven, 2007) for a recent overview.

An obvious case where context sensitiveness is useful is when the user wants to retrieve something closely related to the current situation, such as contact information of a conversation partner. Context can also substitute typed queries when the user does not remember the correct keywords or when the ability to type search terms is limited because of restrictions of mobile devices. Even when the conventional query is available, it can also be helpful to restrict the search space and thus allow the search engine to return correct results with fewer iterations.

2.4 Implicit Speech-based Input

An interface could monitor the user's speech or discussion in order to extract additional information about the user's interest by using automatic speech recognition. Speech recognition in human-computer interfaces has been a subject of extensive study (for a review, see Rebman et al (2003)). For example, Hahn et al (2005) has augmented the observed sound information by a model of attention based on measuring the head posture, Cohen et al (1997) by recognizing the gestures, and Sun et al (2008) by measuring gaze on a computer display. However, as far as we are aware, the idea of combining gaze-based and speech-based implicit input about the interests and context in interaction with persons and objects in the real world is novel.

2.5 Relevance Estimation from Gaze

Studies of eye movements during natural behavior, such as driving a car or playing ball games, have shown that eye movements are highly task-dependent and that the gaze is mostly directed towards objects that are relevant for the task (Hayhoe and Ballard, 2005; Land, 2006). Tanriverdi and Jacob (2000) studied the use of gaze as a measure of interest in virtual environments. Eye tracking has been used as source of implicit feedback for inferring relevance in text (Joachims et al, 2005; Puolamäki et al, 2005), and image (Kozma et al, 2009; Oyekoya and Stentiford, 2006) retrieval applications on a conventional desktop computer. In a recent work, Ajanki et al (2009) constructed implicit queries for textual search from reading patterns on documents. These results indicate that gaze direction is a useful information source for inferring the focus of attention, and that relevance information can be extracted even without any conscious effort from the user. The Virtual Laboratory Guide implements an attentive interface that, in the taxonomy of Vertegaal (2002), monitors the user implicitly and tries to model the behavior of the user in order to reduce his attentive load.

Gaze-controlled augmented reality user interfaces are an emerging research field. So far, the research has been concentrated on the problem of explicitly selecting objects with gaze (Park et al, 2008; Pfeiffer et al, 2008; Nilsson et al, 2009). The conceptual idea of using gaze to monitor the user's interest implicitly in a mobile AR system has been presented previously (Nilsson et al, 2009; Park et al, 2008), but until now a working system has not been demonstrated or evaluated.

Gaze tracking provides an alternative, hands-free means of input entry for a ubiquitous device. One way to give gaze-based input is to make selections by explicit looking. This has been successfully applied for text entry (Ward and MacKay, 2002; Bee and André, 2008; Hyrskykari et al, 2000).

However, giving explicit commands by using gaze is not the best solution in a ubiquitous environment, since it demands the full attention of the user and it suffers from the Midas touch effect: each glance activates an action whether it is intended or not, distracting the user. Hence, in our system, we use a relevance inference engine to infer the user preference implicitly from gaze patterns (see Section 3.4).

2.6 Information Filtering

In their papers Julier et al (2000, 2002) have presented the idea of using real-world context as a search cue in information retrieval, and implemented a system which filters information based on physical location, for selecting what is displayed to the user by means of AR. The main purpose of information filtering is to prioritize and reduce the amount of information presented in order to show only what is relevant to the user.

Already in the Touring Machine (Feiner et al, 1997) there existed a logic to show more information and menu choices for objects that had remained in the center of the user's view for a long enough time. This kind of contextual user feedback is, however, more explicit than implicit by nature. With our gaze tracking hardware we have been able to detect the implicit targets of the user's attention and to use that data in information filtering. As described in the previous section, there have been studies on using gaze as a form of relevance feedback, but to our best knowledge, the current work is the first one to use implicit gaze data for contextual information filtering in an AR setup and to evaluate its usefulness with a user study.

3 Components of Contextual Information Access System

In this section we give an overview of the system and describe its main operational parts.

3.1 System Architecture

We have implemented a pilot software system that can be used in on-line studies of contextual information access with several alternative hardware platforms. The general overview of the system architecture is shown

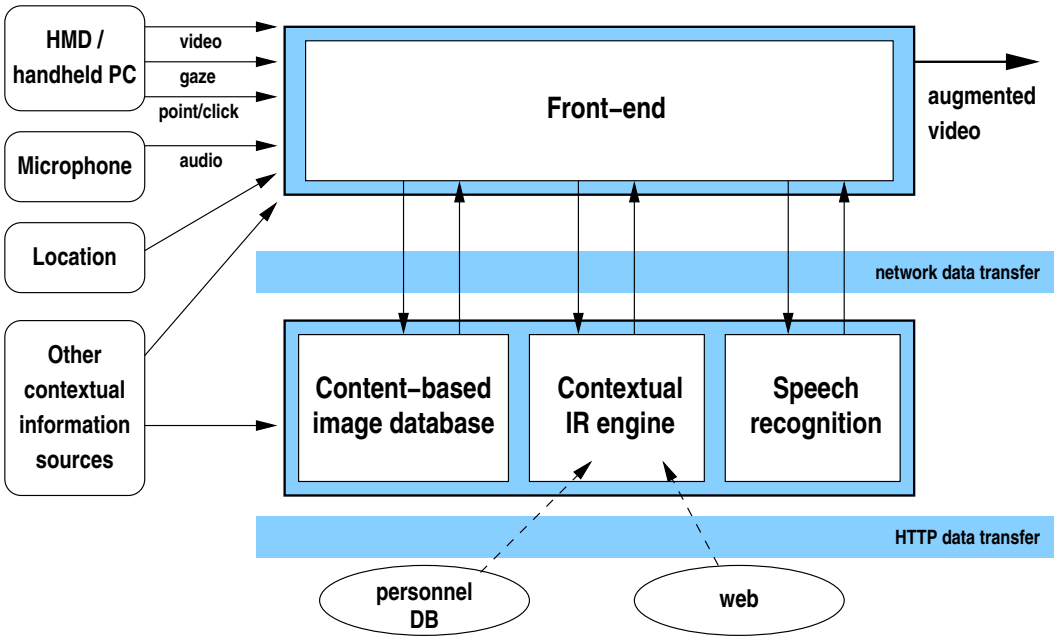

Fig. 2 An overview of the system architecture.

in Figure 2. The technologies integrated into the system include relevance estimation from gaze or pointing patterns; contextual information retrieval; speech, face, object and location recognition; object tracking; and augmented reality rendering.

The front-end user interface of the pilot system is currently operational on two different display hardware platforms: a Sony Vaio ultra-mobile PC (UMPC) and a prototype wearable near-eye camera-see-through display and a miniaturized gaze tracker. See Section 3.6 for a more detailed description of the devices.

All the front-ends communicate wirelessly with common back-end information servers, which are responsible for tasks requiring either more computational resources than what is available on the front-end or dynamic data repositories and access to external data sources. In the current setup, the back-end consists of three separate servers. The first server maintains a context-sensitive database of relevant information to the current application setup and can be queried for context-dependent textual annotations for recognized objects and people. The server also collects background information from external data sources such as the web. The second server is a content-based image and video database engine, which handles image queries and also provides face, object, and scene recognition results. The third server performs speech recognition of the recorded audio, and outputs a stream of recognized words.

3.2 Context-Sensitive Information Retrieval

Our IR database contains annotations for people and objects. Each annotation has an associated context description that is compared to the measured context to decide which annotations are most relevant and should be shown. In the current system the context consist of two types of features: presence features that indicate which objects and people have been seen recently and topic features, which are the estimated relevances of the available topics (such as *research* and *teaching* in the pilot application). Other features, such as time and location, can be added in the future. In the current system the context features of the annotations are fixed, but later they will be inferred and stored on-line.

The current context is inferred from video feed and by observing user actions. The binary presence features are set to one if the corresponding entity has been recognized during the last two minutes, otherwise they are zero. The purpose of the topic relevance features is to track which topics the user prefers. They are periodically adjusted towards the topics of the currently visible annotations. The speech recognizer output is used as additional evidence for topic relevance. If the speech recognizer has been activated during the last two minutes, the topic of the utterance is inferred by comparing the recognized words to topic-specific keyword lists, and the topic feature of the winning topic is set to one and the others to zero.

3.3 Extracting Contextual Cues

The *face recognition* system detects and recognizes human faces and uses them as context cues for associating information to them, such as people's research activities, publications, teaching, office hours and links to their web pages, and Facebook and LinkedIn profiles. Due to real-time performance requirements, recognition is performed in the front-end of the system (see Figure 2) using the OpenCV library. For face detection, we utilize the Viola & Jones face detector (Viola and Jones, 2001), which is based on a cascade of Haar classifiers and AdaBoost. For missed faces (e.g., due to occlusion, changes in lighting, excessive rotation or camera movement), we initiate an optical flow tracking algorithm (Tomasi and Kanade, 1991), which continues to track the approximate location of the face until either the face detector is again able to detect the face or the tracked keypoints become too dispersed and the tracking is lost. The system supports tracking of multiple persons maintaining identities of the detected faces across frames.

The detected faces are transmitted to the image database engine (Figure 2) for recognition using the MPEG-7 Face Recognition descriptor (ISO/IEC, 2002). Before extracting the descriptor, the face images are normalized using region-wise histogram equalization and gamma correction. During tracking, we obtain an increasing number of face images of the same subject that can be used as a set in the recognition stage. Given the set we perform online recognition with a k-nn classifier.

The system can also detect two-dimensional *AR markers* (see Section 3.5) which help in recognizing objects and indoor locations. Markers attached to static objects provide the basis for location recognition. In addition, markers attached to movable objects will be associated with a wider area or a set of places where the objects are likely to be encountered.

The system uses *speech recognition* to gather contextual cues from the user's speech. The recognizer takes speech as an input and gives its transcript hypotheses as an output. The speech transcript is then used to determine the underlying topic of discussion. In the current pilot system, we have a fixed set of potential topics, and the decision between topics is made according to keywords detected from the transcripts.

We use an online large-vocabulary speech recognizer developed at Aalto University (Hirsimäki et al, 2009). The system utilizes triphone Hidden Markov models as context-dependent and gender- and speaker-independent phoneme models trained using sentences selected from the internet and read by a few hundred native Finnish speakers. As a statistical language model, the system has a large 6-gram model of data-driven morpheme-like units.

3.4 Inferring Object Relevance

Real-world scenes typically contain several objects or people, each of which can be considered as a potential source of augmentable information. Distraction and information overload is minimized and best user experience is achieved when information is displayed only for the objects that are interesting for the user at a particular time. For this, a mechanism that estimates the degree of interestingness (relevance) of an object in real time is required.

In the pilot system, the relevance of an object is estimated by the proportion of the total time an object, or related augmented annotations, have been under visual attention within a fixed-length time window. This estimate is defined as *gaze intensity* by Qvarfordt and Zhai (2005). For the HMD, where gaze tracking is available, looking at an object is used as an indicator of attention. For the UMPC, pointing of the display device towards an object is the indicator.

3.5 The Augmented Reality Interface

The objective of the application is to provide the user contextual information non-disruptively. This is accomplished by minimizing the amount of information shown, and by displaying only information that is assumed to be relevant for the user in the given context. The augmentation is designed to be as unobtrusive as possible. All of the visual elements are as simple and minimalistic as feasible and are rendered transparently so that they do not fully occlude other objects.

The user interface of our system is implemented using AR overlay of the virtual objects on top of video of the real world. This is accomplished through a combination of face tracking and 2D marker tracking.

By attaching a marker to an item of interest, we can detect the item's presence and annotate it appropriately (see Figure 3 for an example annotation on a marker).

The AR implementation in the pilot application uses a monocular video see-through display, where augmentations are rendered over the video from a single camera. The camera captures the video with 640×480 resolution at a frame rate of 15 frames per second (FPS). The augmented video is generally displayed at a rate of slightly less than 10 FPS due to the heavy processing involved with respect to the computational power of the mobile devices. We use the ALVAR augmented

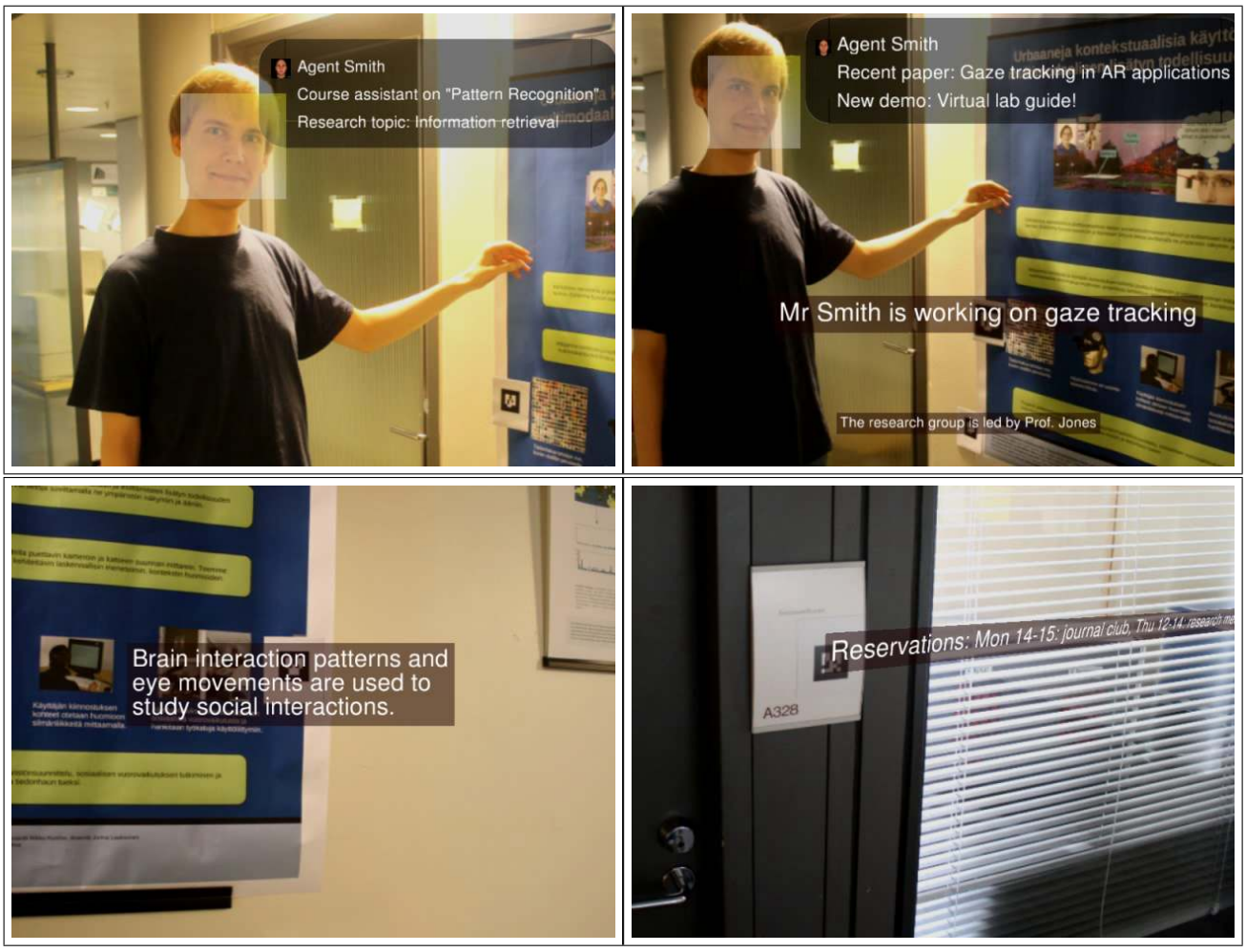

Fig. 3 Screenshots from the Virtual Laboratory Guide. Top left: The guide shows augmented information about a recognized person. Top right: The guide has recognized that the user is interested in research, and shows in the pop-ups mainly research-related information. Bottom left: The user of the guide is looking at a specific section of the poster. Bottom right: The guide is displaying information about reservations to a meeting room.

reality library[3], developed by VTT Technical Research Centre of Finland, for calibrating the camera and for detecting 2D fiducial markers and determining camera pose relative to the markers. For 3D rendering we use the OpenSceneGraph library[4].

The application displays information about people who are recognized using face recognition techniques. Since it is difficult to determine the exact pose of a face relative to a camera, we use a 2.5D approach to place augmented annotations relative to faces. We estimate the distance of a person from the camera based on the size of the detected face in the image and place the augmented information at the corresponding distance facing the viewer. This ensures that the augmentations are consistent with the other depth cues present in the image and helps in associating annotations with persons in the scene. Another cue for association is that the annotation related to a person moves when the person moves. The 2.5D approach is also used in text labels for readability.

In an alternative usage mode the annotations are output as synthetic speech. The synthetic voice for the augmented audio was produced by adapting an average English voice towards the voice of one of our own researchers working in the lab. Using the modelling and adaptation framework (Yamagishi et al, 2009) developed at the EMIME project[5], it is possible to create new personalized voices using just a few samples of the target voice.

[3] http://virtual.vtt.fi/virtual/proj2/multimedia/alvar.html
[4] http://www.openscenegraph.org
[5] `http://emime.org/`

3.6 Display Devices and Cameras

We have implemented alternative device setups to be able to study the effectiveness of the different ways of presenting and interacting with AR content. Ultimately also the question about cost-effectiveness of the setups is important, but in this initial application our focus is on effectiveness.

To show the AR content our system can use two alternative output devices; (1) a near-to-eye display with an integrated gaze tracker and a camera (Figure 4 left), and (2) a handheld UMPC or a standard laptop with a camera (Figure 4 right). The near-to-eye display device is a research prototype provided by Nokia Research Center (Järvenpää and Aaltonen, 2008).

The integrated display and gaze tracker covers partially the field of vision with augmented video of the physical world. The bi-ocular near-to-eye display produces the same virtual image for both of the eyes. Images are perceived larger than the physical display itself and are ergonomic to view because of very low level of geometrical misalignments between the left and right eye images. The device is capable of displaying close to VGA resolution images with a 30 degrees field-of-view in diagonal. The size and weight of the display system are kept reasonably small by expanding the image from a microdisplay with suitable optics, including diffractive optics serving as light guides. Real see-through is not implemented because of the fairly low display brightness and the gaze tracker eye-camera, which is situated just in front of the right eye. The forward-looking camera is situated in between the eyes, both of which are shown the same monocular augmented video output.

The integrated gaze tracker works by illuminating the right eye cornea with collimated infrared beams, which are invisible to human eye. The infrared glints reflected from the pupil are detected by the eye-camera. The locations of the reflections and the pupil center are used for reconstructing the camera-eye-geometry and further, detecting the gaze angle. Estimation of the direction of the visual axis of the eye relative to the head gear can be done after a robust per-user calibration. The tracking accuracy is measured to be less than one degree of visual angle with a speed of 25 measurements per second (Pylvänäinen et al, 2008).

In the second alternative hardware setup, the user carries a small handheld or ultra-mobile Sony Vaio computer with a 4.5 inch display for the augmented video (Figure 4 right). We have replaced the integrated forward-looking video camera with a separate wide-angle web-cam to increase the field of view of the display and to ease the augmentation of several pieces of information on the screen at the same time.

3.7 Interaction

On platforms where the gaze input is unavailable, we utilize explicit pointing as an alternative input method. In the current setting with the UMPC, the user can indicate his or her interest either by pointing the device towards an object or using the stylus to click on an interesting object on the touch screen display. The augmented reality display has a cursor. On the handheld device, the cursor is at the center of the screen, or at the last stylus click location if the stylus has been used less than a second ago. On the head-mounted device the cursor follows the gaze. The potential objects of interest can occlude each other, so we assume that the user is interested in the nearest visible object along the ray defined by the cursor or the gaze.

The gaze and pointing are inputs for the inference engine. The system learns what kind of information is relevant in the current situation by observing which of the objects or the already-shown annotations the user pays attention to. The system shows more information about topics that have recently attracted attention. The shown annotations are considered potential targets of interest for the relevance inference engine, similarly to recognized faces and markers.

The objects and persons may be relevant in different contexts for different reasons. At first, when the system does not yet know what kind of content is relevant, general or uniformly chosen information about the object is shown in the AR annotations.

The estimate of interest we use in the system (based on gaze intensity) is above zero only for the objects that have recently been looked at. We assume that the user is only interested in objects she has looked at, even briefly. This prevents the user from being distracted by annotations that pop up for totally irrelevant objects. The user becomes, to some degree, aware of in advance that an annotation will soon be shown next to the object she is looking at. When the user notices something that is interesting he or she pays more attention to it. The system then shows more information about related topics. If, on the other hand, no attention is paid to the annotations, the system infers that the topics shown are not relevant, and shows less of them in the future. Figure 3 shows example screenshots from the pilot application; in the first screenshot two topics are displayed for the user to choose from, and in the second the user has been inferred to be more interested in research-related annotations.

Fig. 4 The display devices. On the left, the head-mounted display (HMD), a wearable near-to-eye display with integrated gaze tracker. On the right, the ultra-mobile PC (UMPC), a handheld computer with virtual see-through display.

Speech spoken by the user is fed to the speech recognizer. The recognition output is matched to the index terms of the available topic models, and the output of the topic classification is sent back to the system that uses it as a contextual cue.

4 Pilot Application: Virtual Laboratory Guide

As a pilot application for testing the framework we have implemented an AR guide to a visitor in a university department. The *Virtual Laboratory Guide* shows relevant information and helps the visitor to navigate through the department.

The hardware and software platform we are developing is potentially useful in many application scenarios, of which the virtual laboratory guide was chosen as the first one for convenience. The main requirement when setting the system up for a new application is that names or labels for the objects or people of interest need to be given in one way or the other, either when setting up the system or on-line in a user interface. When on-line, either pattern recognition or a marker then identifies the label, and the rest can be done with queries in either a closed (current) or an open (future) system. A sample alternative scenario is virtual tourist guide, where the system could complement the existing tourist guide systems by giving a browser access to unlimited information, chosen according to the observed user interest and context, in addition to the existing fixed information. Another sample scenario is a personal assistant in meetings, helping to remember the names and enabling browsing of background material relevant to discussion context, such as emails from and about the participants.

The current proof-of-concept version of the virtual laboratory guide is implemented on two display devices, namely the head-mounted display with an integrated gaze tracker, and the ultra-mobile Sony Vaio computer. Both devices have a forward-looking video camera and a display that shows the location where the user is looking at, or pointing the computer at.

The task of the system is to infer what information the user would be interested in, and to non-disruptively augment the information in the form of annotations onto the display. See Figure 3 for screenshots of the system in operation.

First the system detects and recognizes the face of a researcher *Agent Smith*, and augments a transparent browser to the display, showing a couple of potentially relevant items about Smith. Based on the user's gaze or pointing pattern the system infers if the user is more interested in Smith's research or courses he teaches, and offers more information retrieved about the appropriate topic. Finally, in the screenshots in the bottom row of Figure 3, the user is focusing her attention to specific markers, and the guide is displaying related information.

The Virtual Laboratory Guide recognizes the context of use and is able to infer the role of the user. Currently only two roles have been explicitly implemented, for concreteness; a student, or a visiting researcher. Later the roles will be inferred implicitly from the contexts that have been activated in the retrieval processes. For a student, the system shows teaching-related information, such as information about the office of a lecturer or a teaching assistant. For a visiting researcher, on the other hand, the guide tells about research conducted in the department. The role is inferred based on which annotated items the user finds interesting (i.e., which annotations he or she looks at or points with the cursor).

The guide needs a database of all recognized persons and objects labelled with markers, and textual

annotations and images associated to them in different contexts. The database can be completely open, even consisting of the whole internet; in the pilot application we use a small constructed database of about 30 persons and objects. The database was constructed manually to keep the setting of the pilot study simple. For people, the database contains their name, research interests, titles of recent publications, and information about lectured courses. For objects, the database contains additional annotations, such as related publications for posters, printer queue names, and names and office hours next to office doors.

5 System Evaluation

5.1 Usability study

A small-scale pilot study was conducted to provide an informal user evaluation of the Virtual Laboratory Guide application, and to test the usefulness of our AR platform. The goal of the study was to collect feedback on how useful people find the ability to access virtual information. We also compared usability of alternative user interfaces: a handheld ultra-mobile PC (UMPC) and a head-mounted (HMD) near-to-eye display, both displaying textual annotations. The head-mounted display is a very early prototype which will naturally affect the results. Figure 4 shows the display devices. The comparisons are designed to tell which kind of augmented reality interface is preferred in information access tasks.

The 8 subjects (all male) were university students or researchers aged from 23 to 32 years old. None of them had prior experience with the Virtual Laboratory Guide application, although some of them had used other AR interfaces before. Each subject used the display configurations in a counterbalanced order to remove order effects. Each subject was trained on how to use the display configurations until they felt comfortable with the technology.

In each condition the subjects were asked to find answers to one research-related question (for example, "Who is funding the research project?") and one teaching-related question (for example, "What is the next exam date for signal processing?"). The answers to the questions were available through the information augmented on the real objects (posters or course material) or on the persons. When the subject had completed one task, the experiment supervisor changed the topic of the augmented information by speaking an utterance containing some keywords for the desired topic to the speech recognizer. The experiment in each condition was stopped after the subject had found the answers to the information search tasks, or after 5 minutes in case the subject was not able to find the answers.

After each display condition the subjects were asked to fill out a subjective survey that had the following six questions:

- How easy was it to use the application?
- How easy was it to see or hear the AR information?
- How useful was the application in helping you learn new information?
- How well do you think you performed in the task?
- How easy was it to remember the information presented?
- How much did you enjoy using the application?

Each question was answered on a Likert scale of 1 to 7, where 1 = *Not very easy/useful/much* and 7 = *Very easy/useful/much*. In addition, subjects were asked what they liked best and least about the display condition and were given the opportunity to write any additional comments about the system.

5.2 Results

In general, users were able to complete the task with either the head-mounted display, or the handheld display and found the system a useful tool for presenting context information. Figure 5 shows a graphical depiction of the questionary results.

In Q2 (how easy was it to see the AR information) the subjects rated the UMPC as easier ($p = 0.016$, Wilcoxon signed rank test). There were no significant differences in the other questions. Most subjects (5 out of 8) preferred the handheld display. In the interview questions the subjects reported that the most severe weakness in the head-mounted display was the quality of the image which made it difficult to read augmented texts. The issues with the quality of the early prototype head-mounted display were to be expected. On the other hand, subjects felt the handheld display had a screen that was too small and the device was too heavy. Two test subjects said that they found the eye tracking and hands-free nature of the head-mounted display to be beneficial. The test subjects found the correct answers in less than 5 minutes with over 90% accuracy in both experiments.

5.3 Follow-up study

In the first experiment we learned that it is difficult to read text on the wearable display because of the limitations of the prototype display technology. Therefore, we decided to make a short follow-up experiment where we

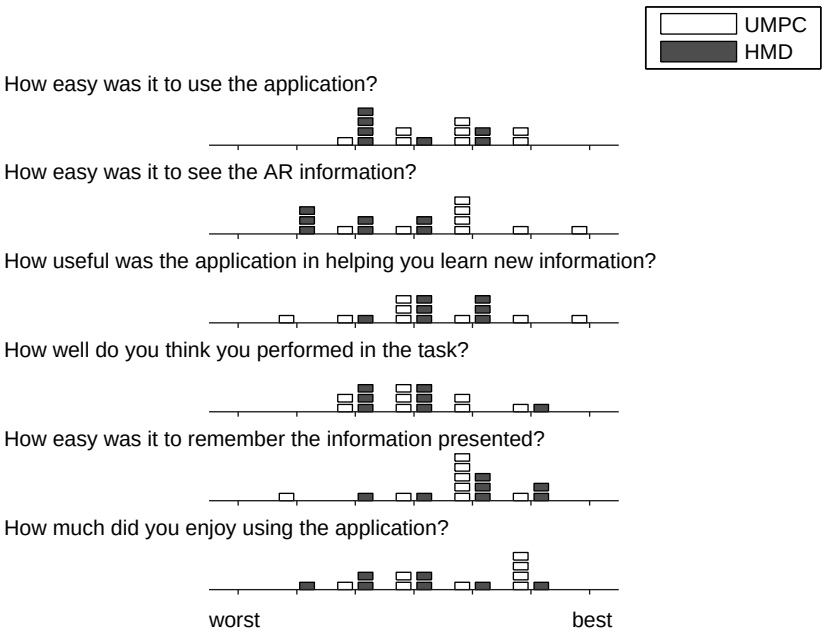

Fig. 5 Results of the usability questionary on 7-point Likert scale in ultra-mobile PC (UMPC) and head-mounted display (HMD) conditions.

collected feedback on using the head-mounted display with the textual annotations replaced by an augmented audio interface.

In the second study, 7 subjects (different from the subjects in the first study, 4 male, 3 female, aged from 25 to 29) completed the same task as in the first study using the head-mounted augmented audio interface. We collected written feedback from the subjects.

5.4 Results of the follow-up study

The users enjoyed the augmented audio interface but still found the head-mounted display uncomfortable to wear. The quality of the synthetic voice got mixed feedback: two test subjects said that the voice was of good quality while two others commented that it was difficult to understand. One test subject noted that it is impossible to skip ahead in speech like one can do while reading. The results of this small-scale study remain inconclusive about the utility of the augmented audio interface in our application. The accuracy of test subjects finding the correct answers was again over 90%.

5.5 Discussion on evaluation results

We have reported results of a preliminary usability studies here. A more thorough evaluations of the system's components is planned. At least studies on how reliably the relevance can be inferred from gaze, what kind of context features are most effective in retrieving relevant information, and how to make the annotations accessible and information easy to browse are needed. A further study is also needed on how relevance inference from gaze and pointing input proposed here compares to other forms of input, such as explicitly selecting the objects or the annotations of interest using a stylus or a 3D mouse, or to standard keyword based queries.

A more formal user study of the whole system will be completed in the future with a second-generation head-mounted display that is more comfortable and has better screen quality and real see-through. However, the results seem to indicate that—unless the head-mounted display has a good image quality—it would be better to provide users with a handheld display they can look at when needed, not all the time.

6 Discussion and Future Work

We have proposed a novel AR application which infers the interests of the user based on her behavior, most notably gaze and speech. The system overlays information about people and objects that it predicts to be useful for the user over the video display of the real world, and uses augmented audio. The system uses implicit feedback and contextual information to select the information to show in a novel pilot application that

combines AR, information retrieval and implicit relevance feedback.

We have implemented our system on two alternate devices: a wearable display with an eye-tracker, and an ultra-portable laptop computer. We have also performed a preliminary evaluation of an application on these hardware platforms. Both devices can be considered as prototype platforms. The current wearable display suffers from fairly low display brightness and virtual see-through, which make using the device somewhat cumbersome. The upcoming releases of the wearable display address these issues. The UMPC, on the other hand, is relatively heavy and likely to remain a niche product. Therefore, we are studying the option of implementing the system on a mobile phone.

In this paper we presented a pilot study on the usability of the idea of augmenting retrieved contextual information onto the view, or as augmented audio. Further studies will be performed on relevance inference, face and marker recognition, speech recognition, as well as the other components.

In the future we will extend the system with generic object recognition capabilities based on techniques such as SIFT features and their matching (Lowe, 1999), which will also reduce the need for markers in object recognition. Also, marker-based AR is limited in that a marker must always be visible in the video for the augmentation to work. This problem can be solved for fixed markers, at least partially, using visual tracking techniques (Davison et al, 2007; Klein and Murray, 2007). Another related problem we will look into is implementing indoor localization.

Output of the speech recognizer will be extended to include multiple recognition hypotheses with confidence measures. The increase of topics from the preliminary two (teaching and research) will enable experiments on on-line language model adaptation. Furthermore, we will improve usability by implementing techniques such as automatic speech detection.

The accuracy of the relevance inference engine will also be improved by extending the current model with additional gaze features and the audio-visual content. Recent experiments (Kandemir et al, 2010) have revealed that a more accurate estimate of relevance can be obtained when additional features of the gaze trajectory, in addition to plain gaze length, are taken into account. Integration of these ideas to the system is currently in progress.

In the visual augmentation, trade-offs need to be made on how much information to show and about how many objects, in order not to occlude the view unnecessarily. In the current system the tradeoff is done purely based on estimated relevance, but later the user needs to be given some direct control about the amount. It is also possible to estimate how often the user actually pays attention to the augmentations, and try to estimate the suitable short-time tradeoff based on that. The balance should at best also take into account the distance and size of the object. Although very small or far-away objects may be equally important to the user, their relevance cannot be estimated equally accurately and it would be a good idea to take into account the uncertainty of the estimated relevance when deciding whether to augment.

Visual augmentation has some additional detailed issues which need to be considered, most notably occlusions and camera geometry. The current system tracks the objects on the scene, and the tracking is tolerant to short-term occlusions. Full long-term occlusions cannot of course be easily modeled at all, but for long-term partial occlusions there is an interface solution: given that the user is estimated to be interested in either the occluding or partially occluded object, both can be shown in the browser.

The present rendering method does not attempt to model the camera geometry in detail — for instance, lens distortion is omitted. It neither matches the augmentations with the illumination conditions, so the augmented objects are easily recognizable from the video. Modelling the camera and illumination with greater accuracy will help to make the augmentation more realistic (Klein and Murray, 2008).

The augmented audio naturally can avoid most of the problems of visual augmentation, at the cost that the audio may be disruptive too in some contexts, and it is harder to infer whether the user pays attention to the audio.

Integration of the enabling technologies in pilot systems will be continued. We plan to use visual location recognition as well as GPS, gyroscopes and accelerometers to provide context information in the future. As we already have a working prototype of the hardware and software framework, new theoretical developments in any of the enabling technologies can be easily integrated into the system and then evaluated empirically in situ. This results from the careful planning of the interoperability and the modular structure of the different software subparts and information servers in the system.

The software and hardware platforms make it possible to test new scenarios or application ideas on a short notice, and study the integration of input modalities, explicit feedback and contextual information retrieval. We are currently in the process of integrating audio output and stereo vision capabilities to the system.

In the pilot study the database was build offline before the experiment. Others have proposed authoring voice (Rekimoto et al, 1998) or 3D object annotations (Baillot et al, 2001) within an augmented reality application and adding them, together with information about the creation context, to the database to be retrieved later. We are planning to extend our system with this kind of interaction capabilities between the user and the virtual world (annotations). For example, we are planning to apply the platform to an architecture-related application, where we overlay architectural design elements on the display of real world, the idea being that an architect can then manipulate and interact with these virtual reality elements.

Acknowledgements Antti Ajanki, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, and Teemu Ruokolainen belong to Adaptive Informatics Research Centre at Aalto University, Antti Ajanki, Melih Kandemir, Samuel Kaski, and Kai Puolamäki to Helsinki Institute for Information Technology HIIT, and Kai Puolamäki to the Finnish Centre of Excellence in Algorithmic Data Analysis. This work has been funded by Aalto MIDE programme (project UI-ART) and in part by Finnish Funding Agency for Technology and Innovation (TEKES) under the project DIEM/MMR and by the PASCAL2 Network of Excellence, ICT 216886.

References

Ajanki A, Hardoon DR, Kaski S, Puolamäki K, Shawe-Taylor J (2009) Can eyes reveal interest? – Implicit queries from gaze patterns. User Modeling and User-Adapted Interaction 19(4):307–339, DOI 10.1007/s11257-009-9066-4

Azuma R (1997) A survey of augmented reality. Presence 6(4):355–385

Baillot Y, Brown D, Julier S (2001) Authoring of physical models using mobile computers. In: Proceedings of the 5th IEEE International Symposium on Wearable Computer, IEEE Computer Society, Washington, DC, USA, pp 39–46

Bee N, André E (2008) Writing with your eye: A dwell time free writing system adapted to the nature of human eye gaze. In: PIT '08: Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer-Verlag, Berlin, Heidelberg, pp 111–122, DOI 10.1007/978-3-540-69369-7_13

Brzezowski S, Dunn CM, Vetter M (1996) Integrated portable system for suspect identification and tracking. In: DePersia AT, Yeager S, Ortiz S (eds) SPIE:Surveillance and Assessment Technologies for Law Enforcement, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, USA, pp 24–35

Cohen PR, Johnston M, McGee D, Oviatt S, Pittman J, Smith I, Chen L, Clow J (1997) QuickSet: Multimodal interaction for simulation set-up and control. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC, pp 20–24

Crestani F, Ruthven I (2007) Introduction to special issue on contextual information retrieval systems. Information Retrieval 10(2):111–113

Davison A, Reid I, Molton N, Stasse O (2007) Monoslam: Real-time single camera slam. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(6):1052–1067, DOI 10.1109/TPAMI.2007.1049

Farringdon J, Oni V (2000) Visual augmented memory (vam). Wearable Computers, IEEE International Symposium 0:167, DOI 10.1109/ISWC.2000.888484

Feiner S, MacIntyre B, Höllerer T, Webster A (1997) A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. Personal and Ubiquitous Computing 1(4):208–217

Hahn D, Beutler F, Hanebeck U (2005) Visual scene augmentation for enhanced human perception. In: International Conference on Informatics in Control, Automation & Robotics (ICINCO 2005), INSTICC Pres, Barcelona, Spain, pp 146–153

Hayhoe M, Ballard D (2005) Eye movements in natural behavior. Trends in Cognitive Sciences 9(4):188–194

Hendricksen K, Indulska J, Rakotonirainy A (2002) Modeling context information in pervasive computing systems. In: Proceedings of the First International Conference on Pervasive Computing, pp 167–180

Henrysson A, Ollila M (2004) Umar: Ubiquitous mobile augmented reality. In: MUM '04: Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia, ACM, New York, NY, USA, pp 41–45, DOI 10.1145/1052380.1052387

Hirsimäki T, Pylkkönen J, Kurimo M (2009) Importance of high-order n-gram models in morph-based speech recognition. IEEE Transactions on Audio, Speech and Language Processing 17(4):724–732

Hyrskykari A, Majaranta P, Aaltonen A, Räihä KJ (2000) Design issues of idict: A gaze-assisted translation aid. In: Proceedings of ETRA 2000, Eye Tracking Research and Applications Symposium, ACM Press, ACM Press, pp 9–14, URL http://www.cs.uta.fi/ curly/publications/ETRA2000-Hyrskykari.pdf

Iordanoglou C, Jonsson K, Kittler J, Matas J (2000) Wearable face recognition aid. In: Proceedings. 2000

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00), pp 2365–2368

ISO/IEC (2002) Information technology - Multimedia content description interface - Part 3: Visual. 15938-3:2002(E)

Järvenpää T, Aaltonen V (2008) Photonics in Multimedia II, Proceedings of SPIE, vol 7001, SPIE, Bellingham, WA, chap Compact near-to-eye display with integrated gaze tracker, pp 700,106–1–700,106–8

Joachims T, Granka L, Pan B, Hembrooke H, Gay G (2005) Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Salvador, Brazil, pp 154–161

Julier S, Lanzagorta M, Baillot Y, Rosenblum L, Feiner S, Hollerer T, Sestito S (2000) Information filtering for mobile augmented reality. In: Augmented Reality, 2000. (ISAR 2000). Proceedings. IEEE and ACM International Symposium on, pp 3 –11, DOI 10.1109/ISAR.2000.880917

Julier S, Baillot Y, Brown D, Lanzagorta M (2002) Information filtering for mobile augmented reality. IEEE Computer Graphics and Applications 22:12–15, DOI http://doi.ieeecomputersociety.org/10.1109/MCG.2002.1028721

Kandemir M, Saarinen VM, Kaski S (2010) Inferring object relevance from gaze in dynamic scenes. In: ETRA 2010: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ACM, New York, NY, USA, pp 105–108

Klein G, Murray D (2007) Parallel tracking and mapping for small AR workspaces. In: Proceedings of Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07), IEEE Computer Society, Washington, DC, USA, pp 1–10

Klein G, Murray D (2008) Compositing for small cameras. In: Proceedings of Seventh IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'08), IEEE Computer Society, Washington, DC, USA, pp 57–60

Kozma L, Klami A, Kaski S (2009) GaZIR: Gaze-based zooming interface for image retrieval. In: Proceedings of 11th Conference on Multimodal Interfaces and The Sixth Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI), ACM, New York, NY, USA, pp 305–312

Land MF (2006) Eye movements and the control of actions in everyday life. Progress in Retinal and Eye Research 25(3):296–324

Lee R, Kwon YJ, Sumiya K (2009) Layer-based media integration for mobile mixed-reality applications. In: International Conference on Next Generation Mobile Applications, Services and Technologies, IEEE Computer Society, Los Alamitos, CA, pp 58–63

Lowe D (1999) Object recognition from local scale-invariant features. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol 2, pp 1150–1157 vol.2, DOI 10.1109/ICCV.1999.790410

Nilsson S, Gustafsson T, Carleberg P (2009) Hands free interaction with virtual information in a real environment: Eye gaze as an interaction tool in an augmented reality system. Psychnology Journal 7(2):175–196

Oyekoya O, Stentiford F (2006) Perceptual image retrieval using eye movements. In: International Workshop on Intelligent Computing in Pattern Analysis/Synthesis, Springer, Xi'an, China, Advances in Machine Vision, Image Processing, and Pattern Analysis, pp 281–289

Park H, Lee S, Choi J (2008) Wearable augmented reality system using gaze interaction. In: Proceedings of the 2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality-Volume 00, IEEE Computer Society, pp 175–176

Pentland A (1998) Wearable intelligence. Exploring Intelligence; Scientific American Presents

Pentland A (2000) Looking at people: Sensing for ubiquitous and wearable computing. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1):107–119, DOI 10.1109/34.824823

Pfeiffer T, Latoschik ME, Wachsmuth I (2008) Evaluation of binocular eye trackers and algorithms for 3D gaze interaction in virtual reality environments. Journal of Virtual Reality and Broadcasting 5(16)

Puolamäki K, Salojärvi J, Savia E, Simola J, Kaski S (2005) Combining eye movements and collaborative filtering for proactive information retrieval. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Salvador, Brazil, pp 146–153

Pylvänäinen T, Järvenpää T, Nummela V (2008) Gaze tracking for near to eye displays. In: Proceedings of the 18th International Conference on Artificial Reality and Telexistence (ICAT 2008), Yokohama, Japan, pp 5–11

Qvarfordt P, Zhai S (2005) Conversing with the user based on eye-gaze patterns. In: CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, New York, NY, USA, pp 221–230, DOI 10.1145/1054972.1055004

Rauhala M, Gunnarsson AS, Henrysson A (2006) A novel interface to sensor networks using hand-

held augmented reality. In: MobileHCI '06: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services, ACM, New York, NY, USA, pp 145–148, DOI 10.1145/1152215.1152245

Rebman CM Jr, Aiken MW, Cegielski CG (2003) Speech recognition in the human-computer interface. Information & Management 40(6):509–519, DOI 10.1016/S0378-7206(02)00067-8

Rekimoto J, Ayatsuka Y, Hayashi K (1998) Augmentable reality: Situated communication through physical and digital spaces. In: Proceedings of the 2nd IEEE International Symposium on Wearable Computers, IEEE Computer Society, Washington, DC, USA, pp 68–75

Singletary BA, Starner TE (2001) Symbiotic interfaces for wearable face recognition. In: Proceedings of HCI International 2001 Workshop On Wearable Computing, New Orleans, LA, pp 813–817

Starner T, Mann S, Rhodes B, Levine J, Healey J, Kirsch D, Picard RW, Pentland A (1997) Augmented reality through wearable computing. Presence: Teleoperators and Virtual Environments 6(4):452–460

Sun Y, Prendinger H, Shi Y, Chen F, Chung V, Ishizuka M (2008) The hinge between input and output: Understanding the multimodal input fusion results in an agent-based multimodal presentation system. In: Conference on Human Factors in Computing Systems (CHI '08), Florence, Italy, pp 3483–3488

Tanriverdi V, Jacob R (2000) Interacting with eye movements in virtual environments. In: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, p 272

Tomasi C, Kanade T (1991) Detection and tracking of point features. Tech. Rep. CMU-CS-91-132, Carnegie Mellon University

Turpin A, Scholer F (2006) User performance versus precision measures for simple search tasks. In: SIGIR '06: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, pp 11–18

Vertegaal R (2002) Designing attentive interfaces. In: Proceedings of the 2002 symposium on Eye tracking research & applications, ACM, p 30

Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01), Kauai, Hawaii, pp 511–518

Wang H, Tan CC, Li Q (2008) Snoogle: A search engine for the physical world. In: Proceedings of the 27th Conference on Computer Communications (IEEE INFOCOM), pp 1382–1390

Ward DJ, MacKay DJC (2002) Fast hands-free writing by gaze direction. Nature 418(6900):838

Yamagishi J, Usabaev B, King S, Watts O, Dines J, Tian J, Hu R, Guan Y, Oura K, Tokuda K, Karhila R, Kurimo M (2009) Thousands of voices for HMM-based speech synthesis. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, ISCA, Brighton, UK

Yap KK, Srinivasan V, Motani M (2005) MAX: Human-centric search of the physical world. In: Proceedings of the 3rd international conference on Embedded networked sensor systems, ACM, pp 166–179