TKK Reports in Information and Computer Science
Espoo 2009 TKK-ICS-R27

UBIQUITOUS CONTEXTUAL INFORMATION ACCESS WITH PROACTIVE RETRIEVAL AND AUGMENTATION

Antti Ajanki, Mark Billinghurst, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki and Timo Tossavainen

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

TKK Reports in Information and Computer Science
Espoo 2009 TKK-ICS-R27

UBIQUITOUS CONTEXTUAL INFORMATION ACCESS WITH PROACTIVE RETRIEVAL AND AUGMENTATION

Antti Ajanki, Mark Billinghurst, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki and Timo Tossavainen

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 5400
FI-02015 TKK
FINLAND
URL: http://ics.tkk.fi
Tel. +358 9 470 01
Fax +358 9 470 23369
E-mail: series@ics.tkk.fi

ISBN 978-952-248-284-6 (Print)
ISBN 978-952-248-285-3 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)
URL: http://lib.tkk.fi/Reports/2009/isbn9789522482853.pdf

TKK ICS
Espoo 2009

ABSTRACT: In this paper we report on a prototype platform for accessing abstract information in real-world pervasive computing environments through Augmented Reality displays. Objects, people, and the environment serve as contextual channels to more information. Adaptive models will infer from eye movement patterns and other implicit feedback signals the interests of users with respect to the environment, and results of proactive context-sensitive information retrieval are augmented onto the view of data glasses or other see-through displays. The augmented information becomes part of the context, and if it is relevant the system detects it and zooms progressively further. In this paper we describe the first use of the platform to develop a pilot application, a virtual laboratory guide, and early evaluation results.

KEYWORDS: Augmented reality, gaze tracking, information retrieval, machine learning, pattern recognition

CONTENTS

1 Introduction 7
1.1 Related Work 8

2 Components of Contextual Information Access System 10
2.1 System architecture 10
2.2 Context-sensitive Information Retrieval 11
2.3 Inferring Object Relevance 12
2.4 Augmented Reality 13
2.5 Display Devices 14
2.6 Interaction 15

3 Pilot Application: Virtual Laboratory Guide 16

4 Evaluation 17
4.1 Results 18

5 Discussion and Future Work 19

References 20

1 INTRODUCTION

In pervasive computing systems there is often a need to provide users with a way to access and search though 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, but 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 Henricksen et. al. point out [10] pervasive computing applications need to place few demands on the user attention and be sensitive to context.

We are interested in the problem of how to efficiently retrieve 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 even be searching, and the information relevant to him 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. Using the implicit measurements about the user's interactions with the environment, we 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 (Fig. 1). The augmented information forms part of the 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 [43]) or image retrieval by zooming deeper into an image collection (Gazir system, [24]).

To realize this, there are several elements needed. Objects and people should be recognized as potential cues with pattern recognition methods. The relevance of them needs to be inferred from gaze patterns and other implicit feedback with machine learning methods. Context-sensitive information retrieval needs to operate proactively given the relevant cues. Information needs to be overlaid on the user's view with AR techniques, on modern display devices. All this should operate such that the user is distracted as little as possible.

We have developed a hardware and software platform which meets these needs and have implemented a first version of a concrete demonstration prototype application showing how it can be applied: this is a *virtual laboratory guide*, which will help a visitor to a university department to find out about research projects and researchers. The virtual laboratory guide presents virtual information about the researchers, offices, and artifacts on see-through browsers that appear as needed and disappear when not attended to.

In the remainder of this paper we first review earlier related work, and describe the lessons learned from this which our research builds on.

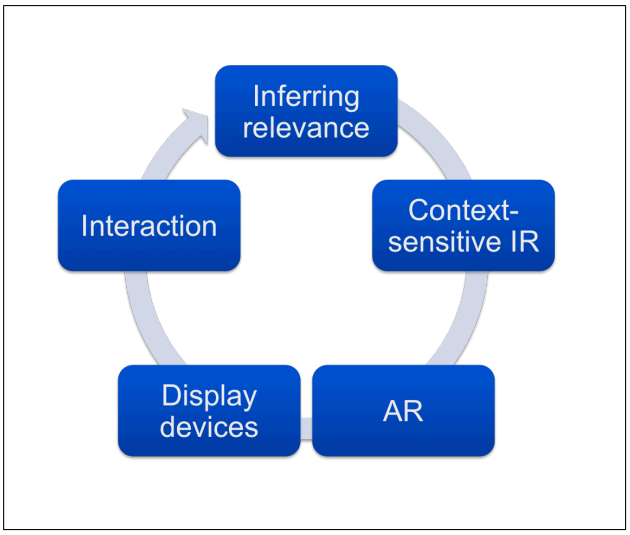

Figure 1: Information retrieval (IR) is done in a loop where relevance of already augmented information (Augmented Reality, AR) and the objects in the scene is inferred from user's observed interaction with them, then more information is retrieved given any contextual cues and the inferred relevance,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

1.1 Related Work

In developing the wearable pervasive information retrieval system described above, our work builds on previous research in the following areas:

Gaze as relevance feedback. We use gaze information for implicit input. Studies of eye movements during natural behavior, such as driving a car or playing ball games, have shown that the eye movements are highly task-dependent and that the gaze is mostly directed towards objects that are relevant for the task [9, 25]. Eye tracking has been used as source of implicit feedback for inferring relevance in text [17, 34] and image [24, 29] retrieval applications on a conventional desktop computer. In a recent work, Ajanki et al. constructed queries for textual search from reading patterns on documents [1]. These results indicate that gaze direction is a useful information source for inferring focus of attention, and that relevance information can be extracted even without any conscious effort from the user.

Speech as relevance feedback. To reveal more about the context and on what properties of the object are particularly interesting, the interface can follow the user's spoken input or discussion. We do this by analysing results of automatic speech recognition. Although speech recognition and human-computer interfaces have been studied a lot (for a review, see [36]), our combination of inferring implicitly both the focus of visual attention from gaze and contextual information from speech is novel.

Contextual information retrieval. Information retrieval (IR) is an established field where the typical research problem is to output documents from a large corpus of documents that best match the given typically textual query. The most visible everyday application of IR is Internet search. The idea in contextual information retrieval is that the retrieval results and how the users interact with the system depends not only of the query but also of the context

of the user. 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 utilizes 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 [5] for a recent overview.

Applications in Mobile Augmented Reality. Augmented Reality (AR) involves the seamless overlay of virtual imagery on the real world [2]. In recent years wearable computers [8] and even cellphones [11] 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. have developed a mobile phone based AR interface with communicates with pervasive sensors to show the temperature distribution of building walls [35].

Previous researchers have used wearable and mobile Augmented Reality systems to display contextual cues about the surrounding environment. For example the Touring Machine [8] added virtual tags to real university buildings showing which departments were in the buildings. These interfaces highlighted the need to filter information according to user's interest, and present it in an uncluttered way. In most cases these interfaces required explicit user input specifying the topics of interest. In our research we wanted to develop a system that used unobtrusive implicit input from the user to present relevant contextual information.

However there has been less research on the use of AR and face recognition to trigger and present contextual cues in a pervasive computing setting. Starner et al. [38] and Pentland [31] describe 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 similarly points out how recognizing people is a key goal in computer vision research and provides an overview of previous work in person identification [32].

Several groups have produced prototype wearable face recognition applications for identifying people [4, 7, 14, 38]. For example Singletary and Starner [37] 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, but they do not go beyond that. 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.

Use of speech recognition in human-computer interfaces has been studied a lot. As far as we know, there have not been studies about integrating gaze-based and speech-based implicit input about the interests and context.

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 [27, 30, 33]. Until now there has been no mobile AR system which has used gaze as implicit input for showing contextual information.

Object location search engines [42, 44] let the user search for physical

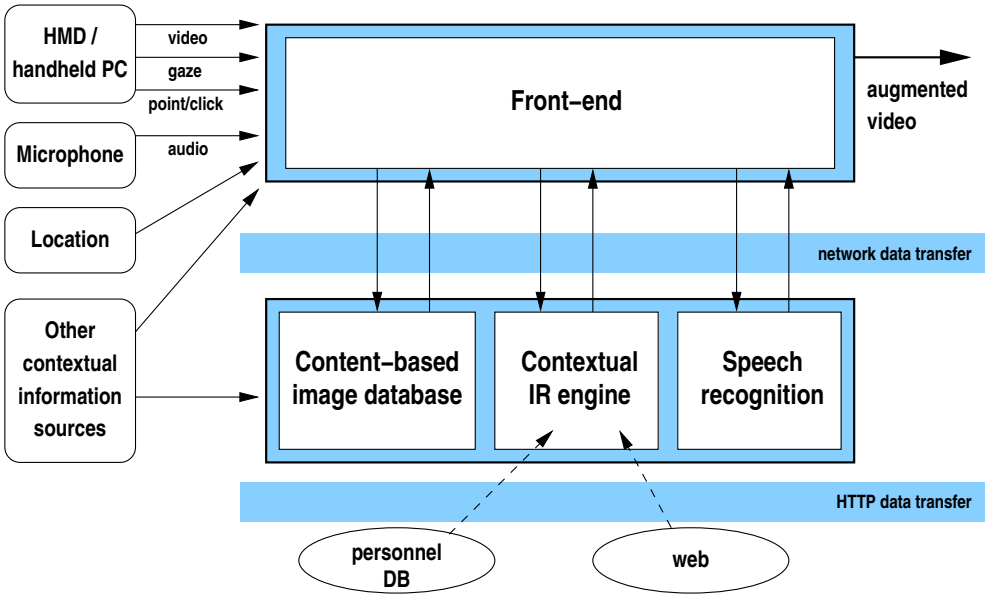

Figure 2: An overview of the system architecture.

location of objects that have been tagged with transponders. However, they require that the user explicitly enters the object ID or terms describing the object to the search engine.

In contrast to this earlier research, our work makes the following research contributions: This is the first work proposing using real world context as a query in information retrieval. There have been studies about gaze as relevance feedback, but we are the first to use gaze in information filtering in an AR setup. The concept of a conceptual information retrieval system using AR is new. We carry out a small-scale user study of the concept.

2 COMPONENTS OF CONTEXTUAL INFORMATION ACCESS SYSTEM

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

2.1 System architecture

We have implemented a pilot software system that can be used in on-line studies of contextual information access with a few alternative hardware platforms. See Figure 2 for a general overview of the system architecture. The integrated technologies 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 operational on two different hardware platforms: a Sony Vaio ultra-mobile PC (UMPC) and in a system equipped with a prototype wearable near-eye see-through display and a miniaturized gaze tracker (see Section 2.5).

All the front-ends communicate wirelessly with common back-end information servers, which are responsible for tasks requiring more computational

resources and dynamic data repositories. In the current setup, one of the servers is a content-based image and video database engine, which handles image queries and also provides face, object, and scene recognition results. The second server can be queried for context-dependent textual annotations for recognized objects and faces. The third server performs speech recognition of the recorded audio, and outputs a stream of recognized words.

2.2 Context-sensitive Information Retrieval

The underlying idea of contextual information retrieval is that a proactive information retrieval and presentation system will be able to deduce from contextual clues what kind of information will be most relevant to the user in his or her current situation. Such contextual clues can be derived from the location, actions and speech of the user and his or her interaction with devices, persons, landmarks, objects, images and sounds in both the true and augmented reality of his vicinity. When combined with implicit and explicit relevance feedback, contextual data cues will provide major improvements and increased comfort in proactive information retrieval compared to conventional explicit search methods.

The context can potentially complement or replace typed search queries. An obvious case where it is useful is when the user wants to retrieve something closely related to the current situation, such as contact information of a conversation partner. If the system is able to correctly predict the information needed, the result is fetched faster and with less effort than by using a conventional search engine. 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 it is possible to perform queries on a conventional search engine, context may be used to restrict the search space and thus allow the search engine to return correct results with fewer iterations.

Our current system supports detection and recognition of human faces and association of various types of information with the recognized persons. This information may be about the person's research activities, publications, teaching, office hours and links to his or her external web pages, including personal homepages, Facebook and LinkedIn.

In the developed system, the recognition of people begins with detecting and tracking of faces. Due to real-time performance requirements, this step is performed in the front-end of the system (see Fig. 2) using the OpenCV library. For face detection, we utilize the Viola & Jones face detector [40], which is based on a cascade of Haar classifiers and AdaBoost. For missed faces (e.g. due to occlusion, changes in lighting, or excessive rotation), we initiate an optical flow tracking algorithm [39], 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 as the identities of the detected faces are maintained across frames.

The detected faces are transmitted to the image database engine (Fig. 2) for recognition using the MPEG-7 Face Recognition descriptor [15]. During tracking we obtain a number of face images of the same subject that can

be used as a set in the recognition stage. Given the set we do the recognition with a Self-Organizing Map -based online multi-example recognition method [23].

The system can also detect two-dimensional AR markers (see Section 2.4) that help in recognition of objects and indoor locations. For location recognition, 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. We plan to use GPS, gyroscopes and accelerometers to provide information about user's position, orientation, and movement in the future.

In the current pilot application, a set of pre-determined objects and locations are detected using markers attached to various places within the department. Such objects and locations include research posters, office rooms, and other objects in the corridors and meeting rooms.

Our system uses speech recognition to collect contextual cues from the user's speech. The speech recognizer of the system turns the speech into a stream of text, or N-best alternative recognition hypothesis, that is utilized for determining the topic of discussion. The definition of topics is based on the frequency of index terms which are either extracted automatically from the available text collection related to the topic or specified manually. The speech recognition and detection of discussion topics are designed in a way that enables both manually specified topics and a completely unsupervised approach where the topics are specified by first clustering all the available text material.

We use an an online large-vocabulary speech recognizer [12], which we in the pilot system limit to output only the best recognition hypothesis. The discussion topic recognition task was also first greatly simplified to choose between the target person's research and teaching interests based on the detection of manually specified keywords from the recognition output. However, the extension to a more general pilot system with more topics and automatically generated index terms is straightforward although it may reduce the accuracy of the topic recognition.

The database from which the contextual IR engine retrieves annotations contains several annotations for the objects and people, and information about what kinds of contexts each annotation would be useful in. The context here is encoded as a feature vector describing the identities of the objects and people present, earlier annotations the user has paid attention to, and keywords detected from speech. Retrieval is done by extracting a feature vector from the current context and history and ranking the contexts in the database based on their similarity to it.

2.3 Inferring Object Relevance

One of the key challenges in the system is real-time identification of which objects or people the user is interested in or needs information about. Each object in the scene can be considered as a potential source of additional information for the user. However, it is necessary to rank these sources by their relevance to avoid information overload.

The user can indicate interesting objects by viewing them, if gaze tracking is available, or by pointing the display device towards them, if not. In the

current pilot system, we assume the relevance of an object to be proportional to the total time an object or related augmented annotations were under user's attention, within a fixed-length time window. The relevance ${r_j}^{(t)}$ of object j at time t is

$$ {r_j}^{(t)} = \frac{{f_j}^{(t-W,t-1)}}{\sum_{i \in V^{(t)}} {f_i}^{(t-W,t-1)}}, \tag{1} $$

where $V^{(t)}$ is the set of detected scene objects and ${f_j}^{(t-W,t-1)}$ is the total duration the user viewed object j within the time window from $t - W$ to $t - 1$.

Based on recent experiments [18] the accuracy of relevance prediction can be improved by taking into account more features of the gaze trajectory than only the total fixation duration. This and further developments will be integrated in the next version of the pilot system.

2.4 Augmented Reality

The user interface of our system is implemented using AR technology: Virtual objects are overlaid on the real world using computer graphics. Usually the virtual objects are rendered on top of video. Realistic rendering of virtual objects requires simulation of the image formation process. This in turn requires a computational model of the camera and the camera's position and orientation (pose) with respect to the real world. Once we have these, 3D computer graphics methods can simulate the camera's image formation from a virtual object. In practice, a pinhole camera with lens distortion correction serves as the camera model and computer vision methods determine the position of a camera with respect to some object or objects of interest. A model for a given camera can be obtained by finding the pinhole model parameters through calibration.

In an uninstrumented environment it is quite difficult to determine the relationship of the camera to the surroundings. Usually this problem is solved by adding fiducial markers to the areas of interest. Fiducial markers are specifically designed for easy detection and pose determination. Obviously, determining the camera pose in some external coordinate system requires that the marker pose in that coordinate system is defined in the application, so that the camera pose can be inferred from the detected marker pose. However, many interesting applications can be built solely on information about the presence of markers and their relative pose to the camera.

By attaching a marker to an item of interest, we can detect the item's presence and annotate it appropriately. Possible annotations include for example 3D models, pictures, and text (see Fig. 4). Tagging all items of interest with fiducial markers is impractical for many applications. In some cases computer vision methods can be used to detect more generic items of interest and their poses. The results for more generic item recognition can be less reliable and may provide partial pose information only. Another major design constraint in practical AR applications is that generally augmented objects can occlude the real world, but the real world cannot occlude the virtual objects, because implementing this would require knowledge of the real world geometry.

The AR implementation in the pilot application is monocular video see-through, where augmentations are rendered on top of of video from a single camera. The camera captures 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 reality library [41] for calibrating the camera and for detecting fiducial markers and determining their pose relative to the camera. The markers are printed on paper and consist of black and white rectangles containing identifying patterns. ALVAR's operation is similar to the well-known ARToolkit [19, 20], but the underlying computer vision algorithms are different. For 3D rendering we use the OpenSceneGraph library [28].

The objective of the pilot application was to provide contextual information non-disruptively. This is accomplished first by minimizing the amount of information shown, and 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. The visual elements are rendered transparently so that they do not fully occlude other objects. We have considered using methods to place the annotations dynamically so that they do not occlude each other, but these are not yet in use in the present version of the system.

The pilot application deals with information related to people working in a laboratory, so in addition to markers we use people as contextual cues. Instead of attaching a marker to every person of interest, we detect people using face recognition techniques. Since it is difficult determine the exact pose of a face relative to a camera, we use a 2.5D approach to augment 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 to 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. The 2.5D approach is also used in text labels for readability. Another cue for association is that an annotation related to a person moves when the person moves. We can also add a connecting line from the annotation to the person, and have different colors in annotations, so a detected face can be highlighted with the same color as the related annotation.

2.5 Display Devices

In a contextual information access system, the information display should be unobtrusive and mobile. Our system can use two alternative output devices; (1) a head-mounted display with an integrated gaze tracker, and (2) a handheld UMPC or a standard laptop with a camera. The head-mounted display device is a research prototype provided by Nokia Research Center [16].

The integrated display and gaze tracker covers the field of vision with augmented video of the physical world (Fig. 3 left). The near-to-eye display produces an image that is perceived to be larger than the physical display itself. The size and weight of the display can be kept reasonably small by expanding the image with suitable optics. The device is capable of displaying VGA

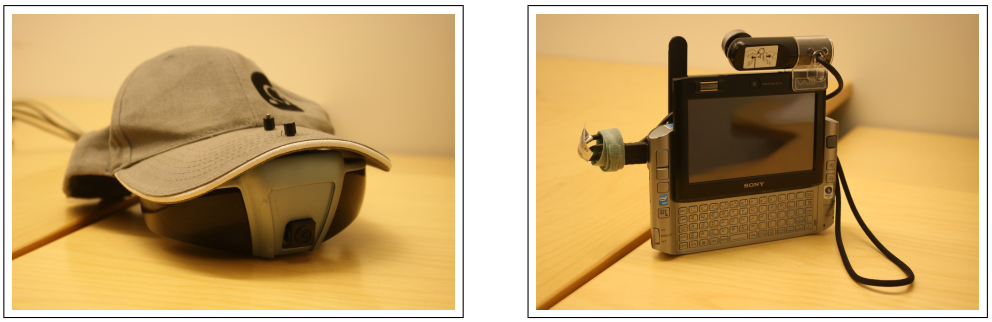

Figure 3: The display devices. On the left, wearable near-to-eye display with integrated gaze tracker. On the right, hand-held computer with virtual see-through display.

quality images with a 30 degree field of view. The forward-looking camera is situated in between the eyes, both of which are shown the same augmented video output.

The integrated gaze tracker works by illuminating the cornea with an infrared beam, which is invisible to human eye, and detecting the infrared glints reflected from the pupil using a camera. The locations of the reflections are used to estimate the direction of the optical axis of the eye relative to the head gear. The tracking accuracy is about one degree of visual angle with 25 samples per second.

In the second alternative hardware setup, the user carries a small hand-held or ultra-mobile Sony Vaio computer with an integrated video camera looking in the opposite direction of the 4.5 inch display (Fig. 3 right). The augmented virtual see-through video is shown on the computer screen.

2.6 Interaction

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 [3, 13, 43].

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 the relevance inference engine to infer the user preference implicitly from gaze patterns (see Section 2.3).

The user can indicate his or her interest by looking or by pointing the UMPC towards an object. The gaze and pointing are inputs for the inference engine. For easing the act of indicating an object, the augmented video is superimposed with a crosshair pattern located at the gaze location or in the center of the camera view, when the gaze is unavailable.

The system learns what kind of information is relevant in the current situation by observing which of the already shown annotation the user pays attention to. More information about topics that have attracted attention previously is show in the future. The 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 random information about the object is shown in AR annotations. 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 them less in the future. Figure 4 shows two screenshots; in the first one, two topics are displayed for the user to choose from, and in the second one the user has been inferred to be more interested in research-related annotations.

3 PILOT APPLICATION: VIRTUAL LABORATORY GUIDE

The hardware and software platform that we have developed would be useful in many application setups. As a pilot application for testing the framework we have implemented an AR guide to a visitor at a university department. The Virtual Laboratory Guide shows relevant information and helps the visitor to navigate the department.

The guide is currently 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 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 augment the information in the form of annotations onto the display non-disruptively. See Figure 4 for a screenshot of the system in operation. First the system detects and recognizes the face of researcher Agent Smith, and augments a transparent browser to the display, showing a couple of generally relevant items about Smith. Based on the user's gaze or pointing pattern the system infers that the user is more interested in research than courses, and offers more information retrieved about the research of Agent Smith.

The Virtual Laboratory Guide recognizes the context of use and is able to infer the role of the user. Currently for concreteness two roles have been explictly implemented; later the roles will mainly be implicit in the sense of being induced through the contexts that have been activated in the retrieval processes. The first pilot role is a student, for whom the system assumes that teaching-related information, such as information about the office of a lecturer or a teaching assistant, is more useful. 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 points with the crosshair).

The guide has a database of all recognized persons and objects labelled with markers and short textual annotations and images associated to them in different contexts. In the pilot application we use a small manually constructed database of about 30 persons and objects. For people, the database contains 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,

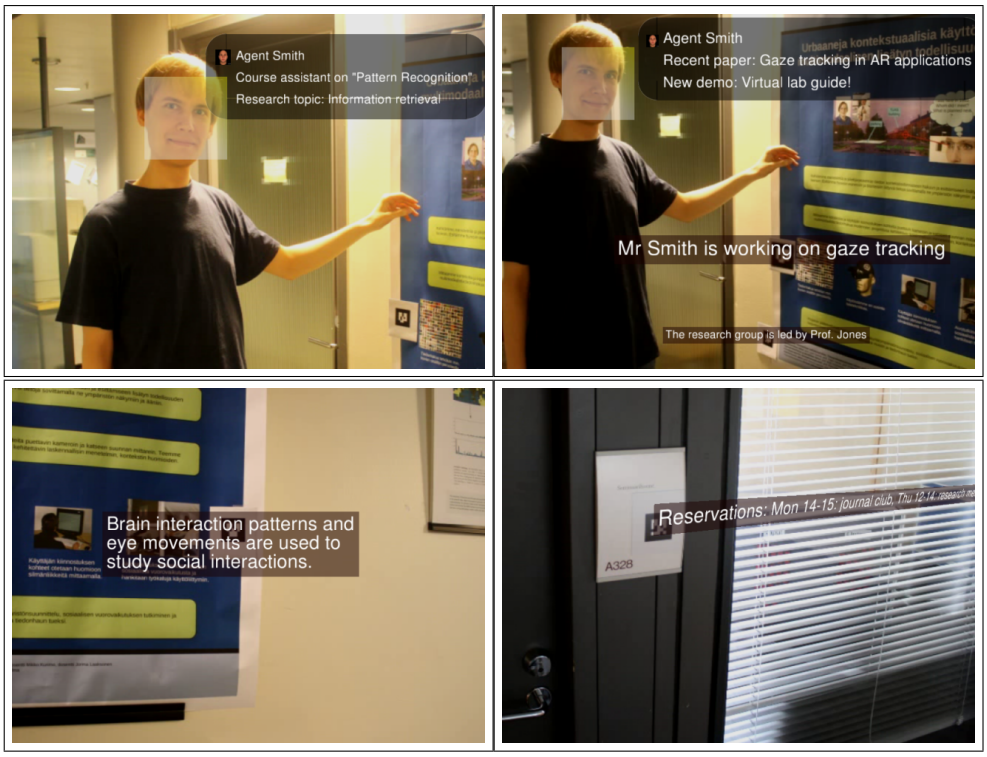

Figure 4: Screenshots from the Virtual Laboratory Guide. Top left: The guide shows augmented information about a recognized person. Top right: The guide has recognized that the user is interested in research-related information. Bottom left: The user of the guide is looking at a specific section of the poster. Bottom right: The guide is displaying information about reservations to a meeting room.

and names and office hours next to office doors.

4 EVALUATION

A small-scale pilot study was conducted to provide an informal user evaluation of the Virtual Laboratory Guide application. The goal of the study was to collect user response to the system and to compare user feedback to using the headmounted display, or the handheld display. In this first pilot study, we compare the usability of the two alternative user interfaces. The head-mounted display is a very early prototype which will naturally affect the results, but we carry out a comparison to get more focused user feedback. To keep the setup as simple as possible we do not use the gaze tracking or speech recognition. However, face recognition, marker detection and contextual image retrieval, which are the key components of the system, have been tested.

Six subjects used the Virtual Laboratory Guide application in two different configurations:

1. HHD: a handheld display on Sony Vaio UMPC

2. HMD: a head-mounted display

In the HMD condition the headmounted display was connected to a laptop computer that was carried in a bag by the subject. Each subject used both display configurations in a counterbalanced order to remove order effects.

The subjects were university students (2 female and 4 male) aged 24 to 30 years old. None of them had experience with the Virtual Laboratory Guide application, although some had used other augmented reality systems in the past.

Before the experiment started each subject was trained how to use both display configurations and spent several minutes using each until they felt comfortable with the technology.

In each display condition the subjects then completed a navigation and information gathering task. Two rooms had fiducial markers in them attached to objects of interest such as academic posters or staff desks. The subjects were told to go to one of the rooms and spend at most 5 minutes looking at all the AR markers to read and learn all the virtual information tags. While the subject was in the room an experimenter entered and they needed to use the face recognition capability of the system to find out who the person was and what his/her key research interests were. Each subject only explored one room per condition and the rooms were also counterbalanced across conditions to reduce order effects. Figure 3 shows the HHD and HMD devices.

After each display condition the subjects were asked questions about the information they had read on the virtual tags and the person they saw in the room. They were also asked to fill out a subjective survey that had the following six questions:

Q1: How easy was it to use the application?

Q2: How easy was it to see the AR information?

Q3: How useful was the application in helping you learn new information?

Q4: How well do you think you performed in the task?

Q5: How easy was it to remember the information presented?

Q6: 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 and 7 = Very Easy. In addition subjects were asked what they liked best and least about the display condition and were given the opportunity to write any other comments about the system.

4.1 Results

Figure 5 shows the result for each of the evaluation questions averaged across all subjects. This is just a pilot study and the subject numbers were too few to conduct a rigorous data analysis, but from the graph the users rated the handheld display much higher for the first four questions. In particular they felt it was much easier to see the AR information in the handheld display (Q2) and the handheld display was more useful in helping to learn new information (Q3). Interestingly, despite these differences, the users felt that both display was equally good at helping them remember the information presented (Q5).

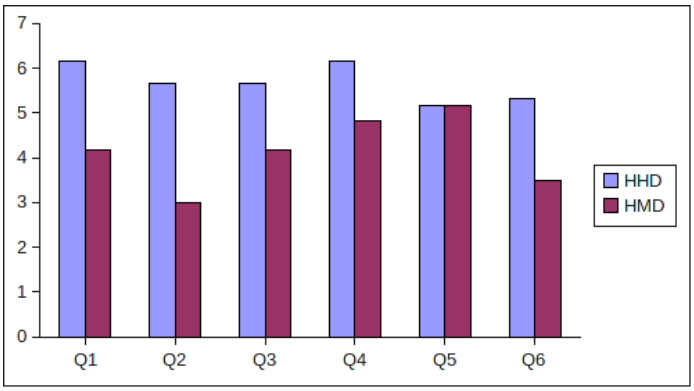

Figure 5: Results of the usability questionary on the Likert scale of 1 to 7 in handheld display (HHD) and headmounted display (HMD) conditions. See the body text for the description of questions Q1–Q6.

The interview questions also supported these results. Most subjects (4 out of 6) said that what they liked best with both types of display was seeing the AR tags on the real world. When asked which display condition they preferred all of the subjects picked the handheld display. Subjects felt the headmounted display was heavy and the image quality blurry and difficult to read the virtual information. One person even complained that the low frame rate on the display gave them motion sickness. In contrast they thought the handheld display was easier to read and had a better display, but they thought the device was heavy and the screen small. Interestingly none of the subjects commented on how the hands-free nature of the head-mounted display was beneficial, or the fact that they had the display in front of their eyes the whole time. Notice that the headmounted display is an early prototype and the issues with the prototype perceived by the users were to be expected.

This is just a preliminary study. A more formal user study of the whole system and thorough evaluation of its components will be completed in the coming months with a new headmounted display that is more comfortable. However the results seem to indicate that unless the head worn display has good image quality and frame rate then it would be better to provide users with a handheld display that they can look at when needed, not all the time.

5 DISCUSSION AND FUTURE WORK

We have proposed a novel AR reality application which infers the interests of the user based on the behaviour of the user, most notably gaze and speech. The system overlays information about people and objects that it predicts to be useful for the user over the display of the real world. It 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 and eyetracker, and an ultra-portable laptop computer, as well as completed a preliminary evaluation of these hardware platforms.

In this paper we presented a pilot study comparing two user interfaces. Further studies will be performed on relevance inference, face and marker

recognition, speech recognition, as well as other components. Furthermore, a more detailed study on the usability of the contextual information retrieval in a real application is needed.

In the future we will extend the system with generic object recognition capabilities based on techniques such as SIFT [26], 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 [6, 21]. Another related problem we will look at is including indoor localization.

The present rendering method does not attempt to simulate the camera in detail — for instance, lens distortion is omitted — or otherwise match the augmentations with the illumination conditions, so the augmented objects are easily recognizable from the video. Simulating the camera and illumination with greater accuracy will help to make the augmentation more realistic [22].

Integration of the enabling technologies in pilot systems will be continued. 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 of the system.

The software and hardware platform 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.

Another area of development is the interaction between the user and the virtual world. In the current implementation, there is little explicit interaction between the user and the virtual world (annotations), but in some other applications interaction with the virtual environment is essential. For example, we are planning to apply the platform to an architecture-related application, where we overlay architectural design elements to the display of real world, the idea being that an architect can then manipulate and interact with these elements.

REFERENCES

[1] Antti Ajanki, David R. Hardoon, Samuel Kaski, Kai Puolamäki, and John Shawe-Taylor. Can eyes reveal interest? – Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction*, 19(4):307–339, 2009.

[2] R. Azuma. A survey of augmented reality. *Presence*, 6(4):355–385, 1997.

[3] Nikolaus Bee and Elisabeth André. 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, pages 111–122, Berlin, Heidelberg, 2008. Springer-Verlag.

[4] S. Brzezowski, C. M. Dunn, and M. Vetter. Integrated portable system for suspect identification and tracking. In A. T. DePersia, S. Yeager, and S. Ortiz, editors, *SPIE: Surveillance and Assessment Technologies for Law Enforcement*. SPIE, 1996.

[5] Fabio Crestani and Ian Ruthven. Introduction to special issue on contextual information retrieval systems. *Information Retrieval*, 10(2):111–113, 2007.

[6] A.J. Davison, I.D. Reid, N.D. Molton, and O. Stasse. MonoSLAM: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, June 2007.

[7] Jonny Farringdon and Vanessa Oni. Visual augmented memory (VAM). In *Proceedings of the 4th IEEE International Symposium on Wearable Computers*, pages 167–168, Los Alamitos, CA, 2000. IEEE Computer Society.

[8] Steven Feiner, Blair MacIntyre, Tobias Höllerer, and Anthony Webster. A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Personal and Ubiquitous Computing*, 1(4):208–217, 1997.

[9] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005.

[10] Karen Hendricksen, Jadwiga Indulska, and Andry Rakotonirainy. Modeling context information in pervasive computing systems. In *Proceedings of the First International Conference on Pervasive Computing*, pages 167–180, Zurich, Switzerland, 2002. Springer.

[11] Anders Henrysson and Mark Ollila. UMAR: Ubiquitous mobile augmented reality. In *MUM '04: Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia*, pages 41–45, New York, NY, 2004. ACM.

[12] Teemu Hirsimäki, Janne Pylkkönen, and Mikko Kurimo. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 17(4):724–732, May 2009.

[13] Aulikki Hyrskykari, Päivi Majaranta, Antti Aaltonen, and Kari-Jouko Räihä. Design issues of idict: A gaze-assisted translation aid. In *Proceedings of ETRA 2000, Eye Tracking Research and Applications Symposium*, pages 9–14. ACM Press, 2000.

[14] C. Iordanoglou, K. Jonsson, J. Kittler, and J. Matas. Wearable face recognition aid. In *ICASSP'00: Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2365–2368, Istanbul, Turkey, 2000. IEEE Computer Society.

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

[16] Toni Järvenpää and Viljakaisa Aaltonen. *Photonics in Multimedia II*, volume 7001 of *Proceedings of SPIE*, chapter Compact near-to-eye display with integrated gaze tracker, pages 700106–1–700106–8. SPIE, Bellingham, WA, 2008.

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, Salvador, Brazil, 2005. ACM.

[18] Melih Kandemir, Veli-Matti Saarinen, and Samuel Kaski. Inferring object relevance from gaze in dynamic scenes. In *Eye Tracking Research and Applications Symposium*, 2010. In review.

[19] Hirokazu Kato and Mark Billinghurst. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In *IWAR '99: Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*, page 85, Washington, DC, 1999. IEEE Computer Society.

[20] Hirokazu Kato and et al. ARToolKit. Available: `http://www.hitl.washington.edu/artoolkit/`. Retrieved 16.10.2009.

[21] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR'07: Proceedings of the Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, Nara, Japan, November 2007.

[22] Georg Klein and David Murray. Compositing for small cameras. In *ISMAR'08: Proceedings of the Seventh IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 57–60, Cambridge, September 2008.

[23] Markus Koskela and Jorma Laaksonen. Semantic annotation of image groups with Self-Organizing Maps. In *CIVR'05: Proceedings of 4th International Conference on Image and Video Retrieval*, pages 518–527, Singapore, July 2005.

[24] László Kozma, Arto Klami, and Samuel Kaski. GaZIR: Gaze-based zooming interface for image retrieval. In *ICMI-MLMI'09: Proceedings of 11th Conference on Multimodal Interfaces and The Sixth Workshop on Machine Learning for Multimodal Interaction*, pages 305–312, Boston, MA, 2009.

[25] Michael F. Land. Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296–324, 2006.

[26] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[27] Susanna Nilsson, Torbjörn Gustafsson, and Per Carleberg. 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, 2009.

[28] Robert Osfield and et al. OpenSceneGraph. Available: `http://www.openscenegraph.org`. Retrieved 12.10.2009.

[29] Oyewole Oyekoya and Fred Stentiford. Perceptual image retrieval using eye movements. In *International Workshop on Intelligent Computing in Pattern Analysis/Synthesis*, Advances in Machine Vision, Image Processing, and Pattern Analysis, pages 281–289, Xi'an, China, 2006. Springer.

[30] Hyung Min Park, Seok Han Lee, and Jong Soo Choi. Wearable augmented reality system using gaze interaction. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 175–176. IEEE Computer Society, 2008.

[31] Alex Pentland. Wearable intelligence. *Exploring Intelligence; Scientific American Presents*, 1998.

[32] Alex Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):107–119, 2000.

[33] Thies Pfeiffer, Marc E. Latoschik, and Ipke Wachsmuth. Evaluation of binocular eye trackers and algorithms for 3D gaze interaction in virtual reality environments. *Journal of Virtual Reality and Broadcasting*, 5(16), 2008.

[34] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola, and Samuel Kaski. 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*, pages 146–153, Salvador, Brazil, 2005. ACM.

[35] Malinda Rauhala, Ann-Sofie Gunnarsson, and Anders Henrysson. A novel interface to sensor networks using handheld augmented reality. In *MobileHCI '06: Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services*, pages 145–148, New York, NY, 2006. ACM.

[36] Carl M. Rebman, Jr., Milam W. Aiken, and Casey G. Cegielski. Speech recognition in the human-computer interface. *Information & Management*, 40(6):509–519, 2003.

[37] Bradley A. Singletary and Thad E. Starner. Symbiotic interfaces for wearable face recognition. In *HCII2001 Workshop On Wearable Computing*, New Orleans, LA, 2001.

[38] Thad Starner, Steve Mann, Bradley Rhodes, Jeffrey Levine, Jennifer Healey, Dana Kirsch, Rosalind W. Picard, and Alex Pentland. Augmented reality through wearable computing. *Presence: Teleoperators and Virtual Environments*, 6(4):452–460, 1997.

[39] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[40] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR'01: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[41] VTT Technical Research Centre of Finland. ALVAR. Available: `http://virtual.vtt.fi/virtual/proj2/multimedia/alvar.html`. Retrieved 12.10.2009.

[42] Haodong Wang, Chiu C. Tan, and Qun Li. Snoogle: A search engine for the physical world. In *IEEE INFOCOM'08: Proceedings of the 27th Conference on Computer Communications*, pages 1382–1390, Phoenix, AZ, 2008. IEEE Computer Society.

[43] D. J. Ward and D. J. C. MacKay. Fast hands-free writing by gaze direction. *Nature*, 418(6900):838, 2002.

[44] Kok-Kiong Yap, Vikram Srinivasan, and Mehul Motani. MAX: Human-centric search of the physical world. In *SenSys'05: Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems*, pages 166–179, San Diego, CA, 2005. ACM.

TKK REPORTS IN INFORMATION AND COMPUTER SCIENCE

TKK-ICS-R17 Tuomas Launiainen
Model checking PSL safety properties. August 2009.

TKK-ICS-R18 Roland Kindermann
Testing a Java Card applet using the LIME Interface Test Bench: A case study.
September 2009.

TKK-ICS-R19 Kalle J. Palomäki, Ulpu Remes, Mikko Kurimo (Eds.)
Studies on Noise Robust Automatic Speech Recognition. September 2009.

TKK-ICS-R20 Kristian Nybo, Juuso Parkkinen, Samuel Kaski
Graph Visualization With Latent Variable Models. September 2009.

TKK-ICS-R21 Sami Hanhijärvi, Kai Puolamäki, Gemma C. Garriga
Multiple Hypothesis Testing in Pattern Discovery. November 2009.

TKK-ICS-R22 Antti E. J. Hyvärinen, Tommi Junttila, Ilkka Niemelä
Partitioning Search Spaces of a Randomized Search. November 2009.

TKK-ICS-R23 Matti Pöllä, Timo Honkela, Teuvo Kohonen
Bibliography of Self-Organizing Map (SOM) Papers: 2002-2005 Addendum.
December 2009.

TKK-ICS-R24 Timo Honkela, Nina Janasik, Krista Lagus, Tiina Lindh-Knuutila, Mika Pantzar, Juha Raitio
Modeling communities of experts. December 2009.

TKK-ICS-R25 Jani Lampinen, Sami Liedes, Kari Kähkönen, Janne Kauttio, Keijo Heljanko
Interface Specification Methods for Software Components. December 2009.

TKK-ICS-R26 Kari Kähkönen
Automated Test Generation for Software Components. December 2009.

ISBN 978-952-248-284-6 (Print)
ISBN 978-952-248-285-3 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)