Vision-Based Location-Awareness in Augmented Reality Applications

Daniel Sonntag
German Research Center for AI
Saarbrücken, Germany
sonntag@dfki.de

Takumi Toyama
German Research Center for AI
Kaiserslautern, Germany
takumi.toyama@dfki.uni-kl.de

ABSTRACT
We present an integral HCI approach that incorporates eye-gaze for location-awareness in real-time. A new augmented reality (AR) system for knowledge-intensive location-based work combines multiple on-body input and output devices: a speech-based dialogue system, a head-mounted AR display (HMD), and a head-mounted eye-tracker. The interaction devices have been selected to augment and improve the navigation on a hospital's premises (outdoors and indoors, figure 1) which shows its potential. We focus on the eye-tracker interaction which provides cues for location-awareness.

ACM Classification Keywords
H.5.2 User Interfaces: Input Devices and Strategies, Natural Language, Graphical HCIs, Prototyping

Author Keywords
Augmented Reality, Navigation, Realtime Interaction

General Terms
Experimentation, Human Factors, Performance

INTRODUCTION

We propose a new multimodal interaction system that can also learn important, task-relevant (visual) information for location-awareness. The multimodal interaction system combines multiple on-body input and output devices: a speech-based dialogue system, a head-mounted AR display, and a head-mounted eye-tracker. The interaction devices have been selected to augment and improve the indoor and/or outdoor task by interpreting navigation cues of users (doctors, patients, or visitors) in hospitals. The user is able to learn individual objects (such as a specific navigation sign or for example the position of a sonography device as a point-of-interest) by looking at it and just saying "this is object x," this is the department's sonography device (in room y)". According to this input (we use automatic speech recognition and eye-gaze provides information about the user's focus of attention) a "cognitive" map can be automatically constructed and queried for location-based real-time information presentation in the HMD. For example, in the medical domain, during examination, the patient's medical record can be shown and the previous sonography findings can be highlighted. We firmly believe that exploiting user-gaze can help alleviate the problem of object identification with mobile eye-tracker focus which, in turn, provides very useful point-of-interests for automatic location-awareness. In fact, location-awareness can play a crucial role for automatically inferring context factors in multimodal interaction and reasoning systems. These potentials should be exploited in a mobile navigation environment and in the context of future, intelligent HCI environments. We hypothesise that object recognition through mobile eye-gaze interpretation has the capacity to eliminate the need for GPS or RFID based indoor and outdoor localisation methods for location-awareness.

LAMDa'13 in conjunction with IUI'13, March 19-22, 2013, Santa Monica, CA, USA.

RELATED WORK

Motivated by previous findings showing the relevance of eye-gaze in multimodal conversational interfaces, we extend the passive input idea to active user input in the AR realm. This also extends the work of using the gaze information to resolve the ambiguities of users speech [10]. In general, eye tracking technology has been used to help automatic language processing, for example to evaluate the role of eye-gaze in multimodal reference resolution [5]; eye tracking technology has been used in intelligent user interfaces (IUIs) more and more frequently. For example, the task of conversing with the user based on eye-gaze patterns [6] introduced a nice idea: the possibility to sense users' interest based on eye-gaze patterns and manage computer information output accordingly. Our motivation (in the learning phase) comes from indications that eye movements during object naming indeed reflect linguistic planning processes [4]. Our approach differentiates mainly in how the eye-gaze is being recorded and interpreted. In our case, we use a mobile, head-mounted system where the objects are interpreted for location-awareness. Every object point to a specific location even if there are two similar devices or navigation signs. Most importantly, the recognised objects are input to a reasoning procedure which infers the location of the user. For example, when we identify specific navigation signs and a sonography device (by interpreting the user's gaze and recognising the focussed objects), we can infer a patient examination task in the sonography examination room. Object recognition is a very powerful clue for a location-aware system because those objects (constantly) referring to a location, e.g., a room, can be turned into precise location markers. This requires a learning step which we also implemented with the help of a speech dialogue system.

SPEECH-BASED POI LEARNING STEP

Over the last several years, the market for speech technology has seen significant developments and powerful commercial off-the-shelf solutions for speech recognition (ASR) or speech synthesis (TTS). For industrial application tasks such medicine discourse and dialogue infrastructures are available [7].

We use a full-fletched conversational interface to learn objects at locations to recognise them. The user comments have an exophoric nature. This means, the named objects refer to the visual, extralinguistic environment. In learning mode, the deictic eye focus gestures are generally understood as pointing gestures that indicate real objects, directions, etc. The dialogical interaction, however, can be enhanced to allow for communicative functions apart from the learning activity. The integrated framework achieves an important objective: it is capable of combining the linguistic dialogue domain with the physical mobile eye-tracker (and HMD combination.)

A reaction and presentation module (REAPR) triggers the context-dependent eye-gaze interpretation software. Other context factors, such as patient and examination context, can be smoothly integrated into the context model. When the user says "that's the specific navigation sign Frauenklinik (women's clinic), department 2.4", we interpret the spoken utterance in combination with the eye-tracker's focus point. In our scenario, we focus on the multimodal dialogue interactions that are directly relevant to the active learning part of the eye-tracker scenario:

1. The user activates the microphone button and starts the ASR. (With the head-mounted mobile eye-tracker, eye gestures can be used.)

2. The user says: "learn a new POI," which issues a respective command in the multimodal interface and the eye-tracker connection.

3. Upon object recognition, REAPR gets informed about a *new* POI and remembers the database instance which is stored in the service backend.

4. The user starts the ASR again.

5. The user says: "that's the specific navigation sign Frauenklinik (women's clinic)" which we fuse into an object image database command now containing the object's classification features and the name of the newly created patient database instance.

In our conception, attention reflected by the eye-gaze is guided by top-down, memory-dependent, and anticipatory mechanisms, such as when looking at a sign to get the right direction. In such a case, attentional salience can be scored with the eye-tracker, and computational heuristics (gaze-points do not significantly deviate from one another, then we assume a fixation on a location-relevant object) help to identify objects in such situations. Attentional salience scores according to the eye-tracker's gaze position can potentially be integrated and optimised.

Figure 2. Specific navigation sign

EYE-TRACKER AND HMD SETUP

Over several decades, researchers investigated a lot in the area of eye tracking and gaze-based interfaces. As a result of recent progress of this research area, a light-weight and compact mobile eye-tracker is available today; it enables us to use gaze as an interface in various scenarios [9, 1].

In our multimodal dialogue system, we use the SMI Eye Tracking Glasses (ETG)[1] in order to recognise which POI object the user, for example the doctor, is looking at in the examination room or the patient is looking at while navigating through the hospital premises. In the HMD scenario, we also obtain further information about the gaze position's POIs in the HMD.

ETG is a binocular eye-tracker, which captures the images of both eyes and computes the gaze position in a scene image (which, in turn, is captured by the scene camera located in the center of the glasses.) In order to obtain accurate gaze positions, the user is required to do a system calibration before using it. The calibration is done by looking at one (or three) point(s) indicated by the system. Brother recently released a product (Airscouter), a new head mounted display whose feature is the transparency of the display. The user can still see the environment through the display. We combined this HMD with the ETG.

OBJECT / SIGN RECOGNITION

In this system, we combine an object recognition framework with the eye-tracking system in order to recognise the navigational POI being examined, i.e., to provide the user with the image-content-based automatic navigation capability.

Perceptual salience of the cues provided by the eye-tracker (focus) is a natural by-product of the sign reading process. The context-dependent interpretation of the eye tracker signal, i.e., the object recognition procedure, works as follows (figure 3):

1. The scene image and the gaze position are obtained from the eye-tracker. Then we:

2. Crop the region of gaze with fixed size of window.

3. Extract local image features from the cropped region. We use SIFT (Scale-Invariant Features Transformation) [3].

4. Execute a nearest neighbor (KNN) search and find the nearest feature vector from the database (to improve the speed of search, we may later use an approximate nearest neighbour search like [2] once distinctive feature are identified

[1] http://eyetracking-glasses.com/

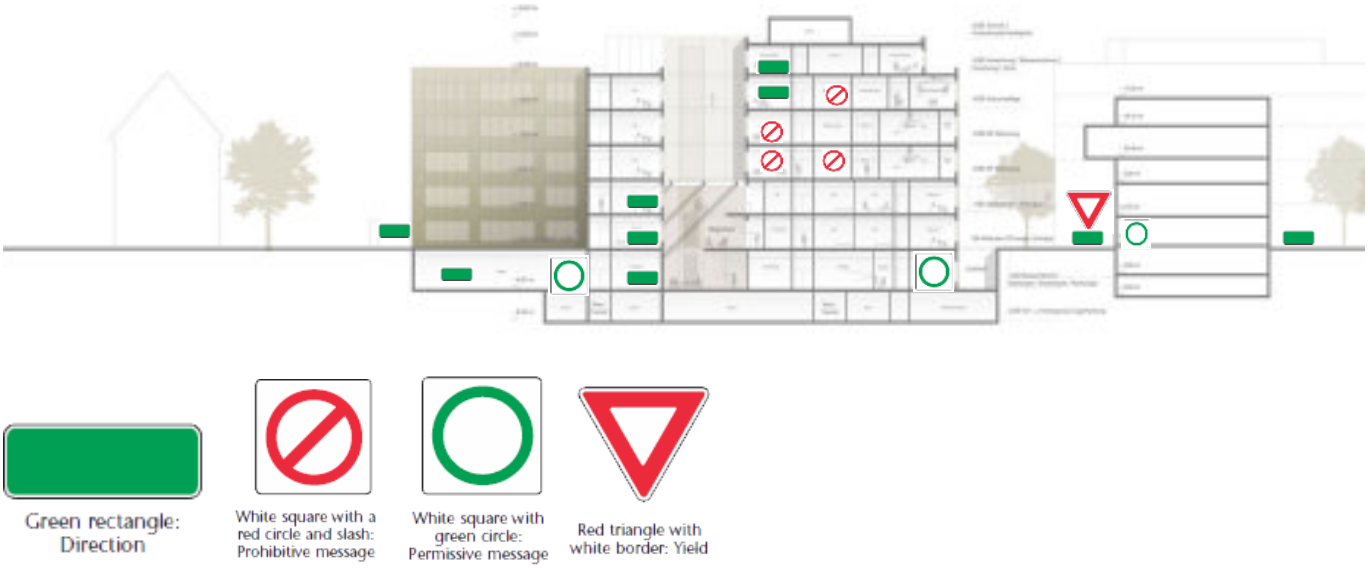

Figure 1. Hospital Premises and Navigation Signs

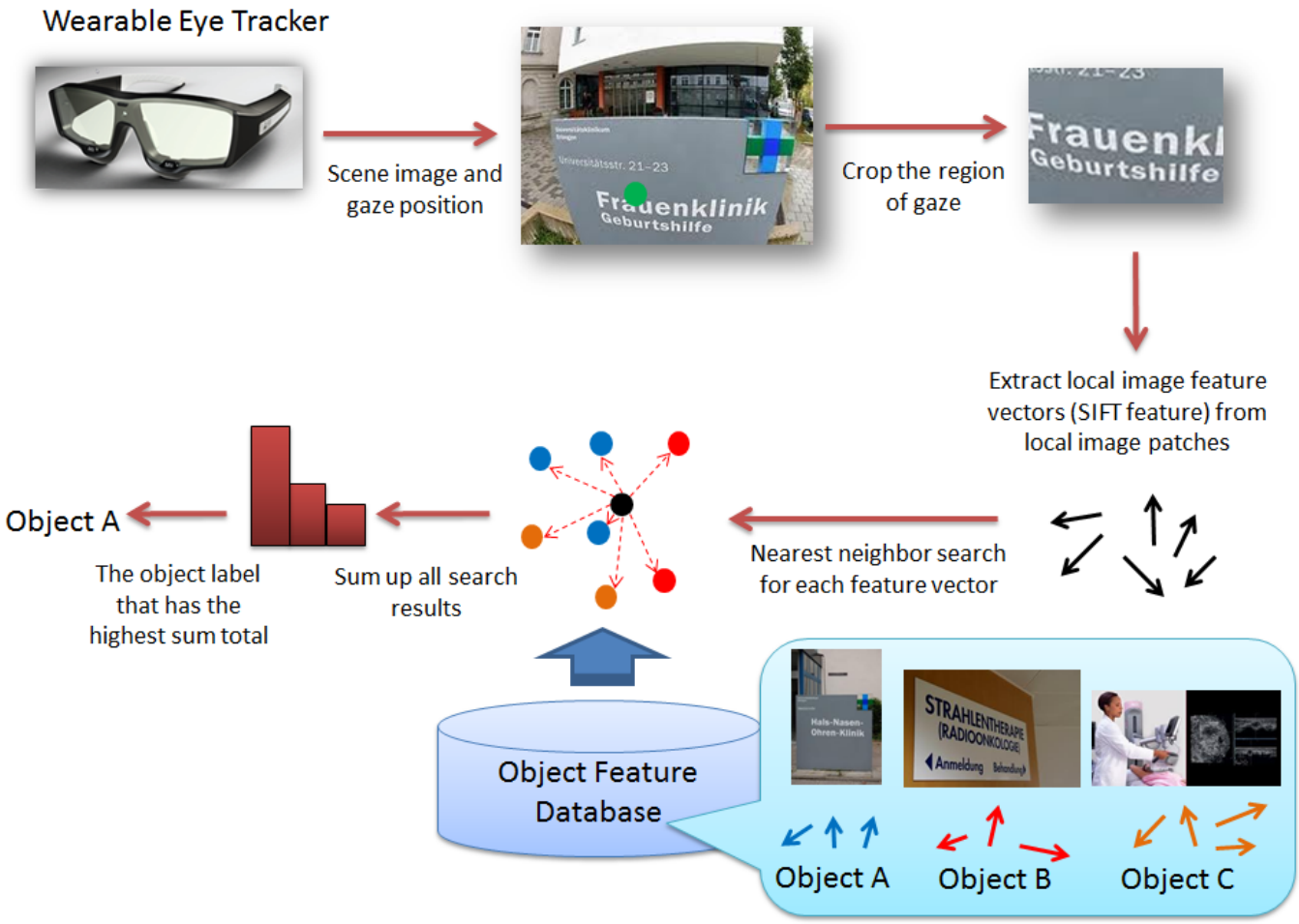

Figure 3. Object / Sign Recognition Process

on a larger population) for each feature vector from the image.

5. Sum up the KNN search results for each feature vector.

6. Return the object label that has the highest sum total as a result (if the sum total is below the threshold value, no result is returned).

7. Send label and further retrieved information to the REAPR display context which triggers the presentation of real-time results in the HMD.

DISCUSSION

There has been a considerable body of work in psycholinguistic studies about the link between eye-gaze and speech, as well as between eye-gaze and internal cognitive processes: it is often said that eye-gaze is a window to the mind. One of he most interesting implications would be that eye-gaze can be used to infer cognitive states and may allow for better mutual grounding theories. As the gaze carries information about the focus of a person's attention, not only navigation-awareness, but even deeper cognitive processes of the user may become visible and interpretable for a AI-based conversational interface or navigation machine. In some way, our on-body head-mounted design with eye-gaze based object/attention recognition paves the way towards machines that "look through the eye of the beholder." Attending to multiple objects at various depths along the direction of the gaze may also obtain additional POIs and concepts for a context-dependent inference step that might help to leverage the simple navigational reasoning of $departmentSign + sonographyDevice \Rightarrow examinationRoom$ or $computerScreen + software + patientFace \Rightarrow patientFindingProcess$. A huge progress could be attributed to future and full generic object character recognition (OCR) rules while "reading" a sign: $Department4, Room24 \Rightarrow location"near"4.24$.

CONCLUSION

In this paper, we have presented an integral approach that incorporates eye-gaze for location-awareness in real-time. In addition, the HMD allows a direct feedback for the user, for example navigation instructions or additional location or situation-based information display. We combined multiple on-body input and output devices, namely a speech-based dialogue system (for the real-time learning scenario), a head-mounted augmented reality display, and a head-mounted eye-tracker, and implemented a specific navigation application context which shows its potential. Our first tests indicate that our mobile gaze-based localization method can provide adequate location awareness when coupled with a tailored interpretation system. Currently we use a simple nearest neighbour search method but this can be extended to an approximate nearest neighbour method such as [2], in order to expand the size of the test database of POIs as location objects to become productive. Our multimodal interaction system combines a mobile eye-tracker with a head-mounted display, and this can now be evaluated in combination with speech-based interaction for its task-based usability [8]. Issues with multiplexed messaging (e.g., poly-social reality), an evaluation of divided attention, and OCR sign reading is subject to future work.

Acknowledgements

This research has been supported in part by the THESEUS Program in the RadSpeech (ERmed) Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. This work was partially supported by EIT ICT Labs.

REFERENCES

1. Bonino, D., Castellina, E., Corno, F., Gale, A., Garbo, A., Purdy, K., and Shi, F. A blueprint for integrated eye-controlled environments. *Universal Access in the Information Society 8*, 4 (2009), 311–321.

2. Indyk, P., and Motwani, R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing* (Dallas, Texas, USA, 1998), 604–613.

3. Lowe, D. Object Recognition from Local Scale-Invariant Features. In *International Conference on Computer Vision* (Kekyra, Greece, September 1999), 1150–1157.

4. Meyer, A. S., Sleiderink, A. M., and Levelt, W. J. Viewing and naming objects: eye movements during noun phrase production. *Cognition 66*, 2 (1998), B25 – B33.

5. Prasov, Z., and Chai, J. Y. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, ACM (New York, NY, USA, 2008), 20–29.

6. Qvarfordt, P., and Zhai, S. Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, ACM (New York, NY, USA, 2005), 221–230.

7. Sonntag, D., Reithinger, N., Herzog, G., and Becker, T. *Proceedings of IWSDS2010—Spoken Dialogue Systems for Ambient Environment*. Springer, LNAI, 2010, ch. A Discourse and Dialogue Infrastructure for Industrial Dissemination, 132–143.

8. Sonntag, D., Weihrauch, C., Jacobs, O., and Porta, D. THESEUS Usability Guidelines for Use Case Applications. Technical report, DFKI and Federal Ministry of Education and Research Germany, 4 2010.

9. Toyama, T., Kieninger, T., Shafait, F., and Dengel, A. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, ACM (New York, NY, USA, 2012), 91–98.

10. Zhang, Q., Imamiya, A., Go, K., and Mao, X. Overriding errors in a speech and gaze multimodal architecture. In *Proceedings of the 9th international conference on Intelligent user interfaces*, IUI '04, ACM (New York, NY, USA, 2004), 346–348.