Listen to What You Look at: Combining an Audio Guide with a Mobile Eye Tracker on the Go

Moayad Mokatren, Tsvi Kuflik and Ilan Shimshoni
The University of Haifa, Mount Carmel, Haifa, 31905
mmokat03@campus.haifa.ac.il, tsvikak@is.haifa.ac.il, ishimshoni@mis.haifa.ac.il

Abstract. The paper presents work in progress about integrating a mobile eye-tracker into a museum visitors' guide system, so to relive the visitor from explicitly requesting information about objects of interest. The novel and most challenging aspects of the study are the image based positioning and the identification of the visitor's focus of attention, while using a commercially available mobile eye tracker. A prototype system has been developed and it will be evaluated in a user study in a realistic setting. The focus of this paper is possible solutions for real-time efficient image based positioning, the overall system design and the planned evaluation.

1 Introduction

Vision is our main sense for gathering information. When we want to gather information about something in our environment, we first look at it. Moreover, when we express interest in something, we look at it. However, the only information we get in this way is what we see: Size, shape, color, distance etc. Nowadays, a lot of additional information about the objects that we see is available online and can be easily accessible when one searches for it. Theoretically, it is available, just a click away, just a query away, or just by activating the mobile device, writing the query, submitting it, scrolling through the results list, selecting the relevant one and accessing the relevant page. This is a bit complicated set of actions in a mobile scenario, when an immediate, personalized and context-aware information is desired. Current technology offers a variety of ways for information delivery to mobile users. Context awareness is the general term describing the attempt to deliver relevant information at the relevant time and place to the user. What is usually common to most context aware services nowadays is that they make use of the communication and computational power (and sensors) of the users' mobile devices (e.g. mostly smartphones). In addition, they interact with their users mainly by their mobile device's touch screens, which has one major limitation: they are limited in size, the users have to look at them during the interaction, and use a keyboard or select icons etc. Even though voice commands can be used for activating applications, this option is still very limited.

A major challenge in the mobile scenario is to know exactly what the user is interested in. In classical human-computer interaction, the users use a pointing device, most commonly a mouse or by touching a touch screen. However, this is becoming a

major challenge in the mobile setting as noted by Calvo and Perugini [2014], who surveyed novel pointing approaches for wearable computing. The user's position is the best hint, accompanied by the user's orientation. Still, there are many possibly interesting objects near and around the user. If we know what the user is looking at, and what the specific user's gazing profile is, then we can narrow down the possibly relevant objects of interest and we can better serve the user with relevant service/information when needed.

As we move towards "Cognition-aware computing" [Bulling and Zander 2014], it becomes clearer that eye-gaze based interaction should and will play a major role in human-computer interaction before/until brain computer interaction methods will become a reality [Bulling et al. 2012]. With the advent of mobile and ubiquitous computing, it is time to explore the potential of mobile eye tracking technology for natural, intelligent interaction of users with their smart environment, not only in specific tasks and uses, but for a more ambitious goal of integrating eye tracking into the process of inferring mobile users' interests and preferences for providing them with relevant services and information, an area that received little attention so far.

Cultural heritage (CH) is a traditional domain for experimentation with novel computing technology. An intelligent mobile museum visitors' guide is a canonical case of a context-aware mobile system. Museum visitors move in the museum, looking for interesting exhibits, and wish to acquire information to deepen their knowledge and satisfy their interests. A smart context-aware mobile guide may provide the visitor with personalized relevant information from the vast amount of content available at the museum, adapted for his or her personal needs. Mokatren et al. [2016] already presented a novel image based positioning technique using mobile eye tracker for a museum visit, where the position of the visitor is identified in a predefined museum layout, and once is determined an object of interest can be inferred. In this work we aim at developing an audio guide system using a mobile eye tracker on the go as a positioning system and as an implicit pointing device for natural interaction with the system using gesture recognition.

2 Background and Related Work

2.1 Requirements for Museum Audio Visitor's Guide

The museum environment has many limitations, such as the restriction not to make noise, not to talk loudly, not to touch anything, etc. It is obvious that museum visitor's mobile guides should not be a replacement for traditional interpretation methods, but rather complement them [Economou, 1998]. Under these limitations Cheverst et al. [2000] have mentioned two key requirements for such guides, the first of which is Flexibility. The system is expected to be sufficiently flexible to enable visitors to explore, and learn about, a museum in their own way, including controlling their own pace of interaction with the system. The second requirement is that the system will be context aware, meaning that the information presented to the visitors should be

tailored to their personal context. The personal context includes, among other things, the visitor's interests, the visitor's current location and exhibits already visited.

2.2 Image Based Positioning

Consider a device consisting of a forward looking camera and an eye tracker. The device takes a picture while the user is fixating on a certain position within the image. The challenge is to recognize the object in the scene in order to deliver content related to this object to the user.

When an image taken by the front camera of the device, it can be matched to a set of existing images, where the goal is to find which of the images shows the same scene as the test image. The matching algorithm should work in cluttered scenes (scenes from which objects have been removed or added), where the images were not taken from the same pose and with varying illumination. For this to work local image features were developed that are unaffected by nearby clutter or partial occlusion. The features are at least partially invariant to illumination, 3D projective transforms, and common object variations. On the other hand, the features must also be sufficiently distinctive to identify specific objects among many alternatives. Several types of local features have been developed. The most popular type of feature is SIFT [Lowe 1999] but others also exist.

Location-awareness procedure using image matching works as follows:

1. A set of images of the exhibits should be taken, each image may contain one or more objects. For each object that appears in an image, a distinct label value and size of region around the object should be given (in terms of width and height – rectangular shape)
2. Eye-tracker scene camera frame is taken and image-to-image matching procedure is applied using SIFT features. The result is an image with labeled regions in the current scene's frame. A pair of images will be marked as matched if the percentage of the matched feature points (as presented by [Lowe 1999]) is larger than some threshold value (the threshold is determined by case study evaluation).
3. Fixation mapping transformation. The fixation point is transformed from the eye tracker scene camera to a suitable/matched region in the image that we got in step one (image from the data-set with labeled regions)

The result of the above procedure is a location id (or an exhibit id in a museum visit) and point/object of interest (specific object in the exhibit that the visitor looked at).

2.3 Pupil-Dev Mobile Eye Tracker

Pupil eye tracker [Kassner et al. 2014] that is presented in Figure 1, is an accessible, affordable, and extensible open source platform for pervasive eye tracking and gaze-based interaction. It comprises a light-weight eye tracking headset, an open source software framework for mobile eye tracking, as well as a graphical user interface to

playback and visualize video and gaze data. Pupil features high-resolution scene and eye cameras for monocular and binocular gaze estimation.

Figure 1. Pupil eye-tracker (http://pupil-labs.com/pupil)

2.4 Mobile Eye Tracker as a Pointing Device

Eye tracking is an active area of research, where significant progress is continuously made over a long time. Recently, Yousefi, et al. [2015] surveyed a large variety of mobile eye tracking applications and technologies, including aviation, marketing, learning, medicine and more. Furthermore, as predicted (and surveyed by [Yousefi, 2015]), relatively inexpensive, easy to use mobile eye trackers are appearing. Usually, they are experimented in specific areas of applications and tasks. Mokatren et al. [2016] proposed a tool for location awareness, interest detection and focus of attention using computer vision techniques and mobile eye-tracking technology, the focus was on a museum visit. The proposed tool is based on image based positioning technique, for that a set of images that represents the layout of the museum should be taken and stored for image to image comparison.

3 Research Goal and Questions

Our goal is to examine the potential of integrating the eye tracking technology as a natural interaction device into mobile audio guide system (e.g. using the eye-tracker as a natural pointing device in a smart environment). Using it as a pointing device that enables systems to reason unobtrusively about the user's focus of attention and suggest relevant information about the focus of attention as needed.

Our focus is on developing a framework for museum's audio guide that extends the work of Mokatren et al. [2016] for information delivery based on eye gaze detection and image based positioning. We will answer the following question: **How can we integrate the mobile eye tracker as a pointing device in a system that delivers audio information to the visitor?**

For that we have developed a prototype of a system that runs on a laptop and uses Pupil Dev [Kassner et al. 2014] mobile eye tracker for identifying objects of interest and delivering informative content to the users.

In our study we have considered different factors and constrains, the real environment lighting conditions (scenes varies in different day time, e.g. direct sunlight, see figure 2 for example) that can greatly affect the process of image based positioning. For that, different dataset images were taken at different times to ensure successful positioning procedure. Another aspect was the position of the objects relative to the eye tracker holder, since the eye tracker device is head-mounted as this is constrained by the environment layout.

Figure 2. Same exhibit at different day time.

4 Context-aware, Mobile Audio Guide Framework

A key challenge in using mobile technology for supporting museum visitors' is figuring out what they are interested in. This may be achieved by tracking where the visitors are and the time they spend there [Yalowitz and Bronnenkant, 2009]. A more challenging aspect is finding out what exactly they are looking at [Falk and Dierking, 2000]. Given todays' mobile devices, we should be able to gain access seamlessly to information of interest, without the need to take pictures or submit queries and look for results, which are the prevailing interaction methods with our mobile devices.

Lanir et al. [2013] discussed the influence of location-aware mobile guide museum visitors' behavior. Their results indicate that visitors' behavior was altered considerably when using a mobile guide. Visitors using a mobile guide visited the museum longer and were attracted to and spent more time at exhibits where they could get information from the guide. Moreover, they argued that "While having many potential benefits, a mobile guide can also have some disadvantages. It may focus the visitor's attention on the mobile device rather than on the museum artifacts".

In this section we describe the implementation of the audio guide framework that will address the above-mentioned two challenges – it will identify users' focus of attention accurately and it will do that unobtrusively. The system uses Pupil Dev [Kassner et al. 2014] mobile eye tracker (as a pinpoint device for inferring object of interest), laptop (that serves as a computational power) and earphones (for audio information delivery). The system extends the image based positioning technique that was presented by Mokatren et al. [2016] to deliver audio information about exhibits in the museum. A visitor wears the mobile eye tracker which is connected to a laptop (carried on back bag) enters the museum, when he looks steadily for approximately three seconds at an exhibit, the image based positioning procedure starts and location/position and point of interest is identified.

We have implemented two versions of audio mobile guide:

1. Reactive: After identifying the position of the visitor and point/object of interest, "beep" sound is played and immediately after that audio information about the exhibit is delivered (see Figure 3).
2. Proactive: After identifying the position of the visitor and point/object of interest, "beep" sound is played, and the system wait for mid-air gestural action (stop sign). After performing the gestural action, audio information is delivered (see Figure 4).

For both systems we have added an option to stop the audio information delivery at any time by performing mid-air gestural action (stop sign).

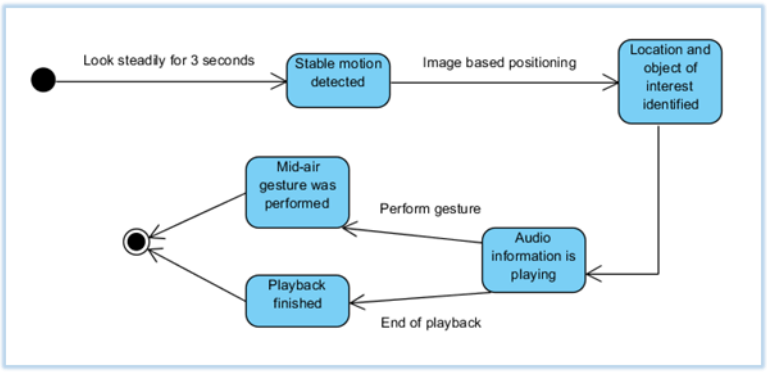
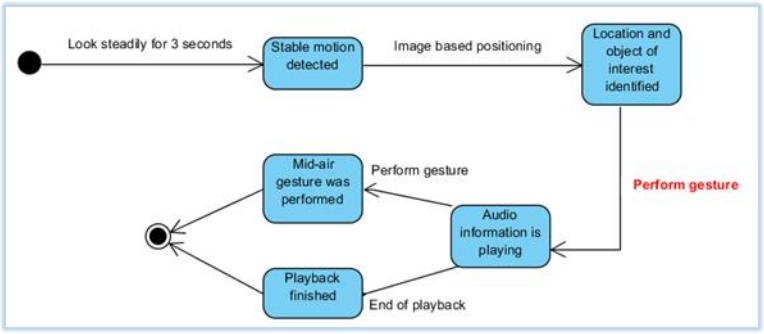


Figure 3. State machine diagram for the **reactive** version



Figre 4. State machine diagram for the **proactive** version

5 Experiment Design

The system will be evaluated in user studies; the participants will be students from University of Haifa. The study will be conducted in Hecht museum[1], which is a small museum, located at the University of Haifa that has both an archeological and art collections.

[1] http://mushecht.haifa.ac.il/

The experiment will be within-subject design that will compare the use of the audio guide with the two versions. The study will include an orientation about using the eye tracker, mid-air gestural interaction (one type of gesture – "stop sign") and the mobile guide, then taking a tour in the museum with the audio guide.

The exhibits will be divided into three categories: Small exhibits, large exhibits and showcases (vitrine shelves). Each case-study will include exhibits from the three categories, we will try to differentiate each case-study exhibits by choosing different exhibits from the same category to reduce as possible the effect of learning.

Data will be collected as follows: The students will be interviewed and asked about their visit experience, and will be asked to fill questionnaires regarding general questions such as if it is the first time that they have visited the museum, their gender and age, and more.

6 Discussion and Conclusions

In the CH setting, visitors' movement in space, time spent, information requested, vocal interaction and orientation were used for inferring users' interest in museum exhibits. Adding eye gaze as additional source may greatly enhance the ability to pinpoint the user's focus of attention and interest (e.g. on products or exhibits), hence improve the ability to model the user and better personalize the service offered to her/him (e.g., exhibit or product information, shopping assistance).

In this paper we presented a framework for a context-aware mobile museum audio guide that uses mobile eye tracking technology for identifying the location of the visitor and inferring his point/object of interest. The audio guide system framework consists of two versions: Reactive and proactive, in the reactive version audio information is delivered immediately once the point of interest is identified, in contrast to the proactive version where the visitor needs to perform a mid-air gestural action to start the audio delivery. The system has not been evaluated yet.

In the image based positioning technique, there is overhead time in matching the camera scene image with every image from the dataset. If the visitor stands at a fixed point and a little time has passed since the last match procedure, then we can search for a matched image from the physical nearest neighbors only. For that we need to represent the data set using a graph, each node will represent the exhibit image/label and the arc value represent the physical distance.

Future work will focus on optimizing the image based positioning procedure by representing the museum layout using graph, and then evaluating the system in an experiment in a museum, in a realistic setting of a museum visit.

References

1. Bulling, A., Dachselt, R., Duchowski, A., Jacob, R., Stellmach, S., & Sundstedt, V. (2012). Gaze interaction in the post-WIMP world. In CHI'12 Extended Abstracts on Human Factors in Computing Systems, 1221-1224. ACM.

2. Bulling, A., & Zander, T. O. (2014). Cognition-aware computing. Pervasive Computing, IEEE,
3. Calvo, A. A., & Perugini, S. (2014). Pointing devices for wearable computers. Advances in Human-Computer Interaction, 2014.
4. Cheverst, K., Davies, N., Mitchell, K., & Friday, A. (2000, August). Experiences of developing and deploying a context-aware tourist guide: the GUIDE project. In Proceedings of the 6th annual international conference on Mobile computing and networking (pp. 20-31). ACM.
5. Economou, M. (1998). The evaluation of museum multimedia applications: lessons from research. Museum Management and Curatorship, 17(2), 173-187.
6. Lowe, David G. "Object recognition from local scale-invariant features." (1999). The proceedings of the seventh IEEE international conference on. Computer vision. Vol. 2 ,pp. 1150-1157
7. Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. 1151-1160. ACM.
8. Lanir, J., Kuflik, T., Dim, E., Wecker, A. J., & Stock, O. (2013). The influence of a location-aware mobile guide on museum visitors' behavior. Interacting with Computers, 25(6), 443-460.
9. Mokatren, M., Kuflik, T. and Shimshoni, I. (2016) Exploring the potential contribution of mobile eye-tracking technology in enhancing the museum visit experience. Accepted to the workshop on Advanced Visual Interfaces in Cultural Heritages – AVI-CH 2016 – co-located with AVI 2016.
10. Yalowitz, S.S. and Bronnenkant, K. (2009) Timing and tracking: unlocking visitor behavior. Visit. Stud., 12, 47–64.
11. Yousefi, M. V., Karan, E. P., Mohammadpour, A., & Asadi, S. (2015). Implementing Eye Tracking Technology in the Construction Process. In 51st ASC Annual International Conference Proceedings.