Museum Guide 2.0 – An Eye-Tracking based Personal Assistant for Museums and Exhibits

Takumi Toyama, Thomas Kieninger, Faisal Shafait, Andreas Dengel
German Research Center for Artificial Intelligence
<firstname.lastname>@dfki.de

Abstract

This paper describes a new prototypical application that is based on a head mounted mobile eye tracker in combination with content based image retrieval technology. The application, named "Museum Guide 2.0", acts like an unintrusive personal guide of a visitor in a museum. When it detects that the user is watching a specific art object, it will provide audio information on that specific object via earphones.

The mobile eye tracker thereby observes the visitors eye movements and synchronizes the images of the scene camera with the detected eye fixations. The built in image retrieval subsystem recognizes which of the art objects in the exhibition is currently fixated by the users eyes (if any).

Challenges that had to be faced during our research are the modifications of the retrieval process utilizing a given fixation for better accuracy, the detection of consciousness when looking at one specific object as trigger event for information delivery and to distinguish from noise (unconscious fixations).

This paper focuses on the application aspect of Museum Guide 2.0. It describes how a database of given art objects is created from scratch and how the runtime application is to be used. We end with a user study that has been conducted to evaluate the acceptance of the system, specifically in contrast to conventional audioplayer based approaches.

1 Introduction

When tourists visit a Museum or a historical site, they need more information about the exhibits they are visiting. This information is usually provided by a trained professional (tourist guide). However, professional guides cannot cater the needs of all tourists. Therefore, automated personal guides are usually provided to visitors in Museums or archaelogical sites to aid them in getting more information about their exhibits of interest. Traditionally, these guides are provided as pre-recorded audio tapes, where the user can select and listen to an audio track corresponding to the exhibit of interest.

Recent advances in the fields of object recognition, augmented reality, and virtual reality have led to the development of many interesting ideas for enhancing user experiences when visiting a Museum. The CINeSPACE project [Santos et al. 2007] aimed at designing and implementing a mobile rich media collaborative information exchange platform. The main idea was to enable users to interact with location-based multimedia contents while navigating a city. Audiovisual information was delivered through a portable low-cost wireless high definition near-to-the-eye display and audio phones. The audio-visual information focused on the cities' culture, history, tourism and art accessed through film heritage played out in a mobile virtual environment.

The EU project AGAMEMNON aims at providing a visitor to a site of historical interest with personalised, information enriched experience through 3G cellphones and at the same time contributing to the preservation of cultural heritage [Agamemnon].

In the recent years, eye tracking and image based object recognition technologies have reached a certain degree of maturity which encouraged us to develop a new application on top of the technological basis of.

Museum Guide 2.0 (also MG2.0 - the "2.0" like to express a next generation of museum guides) is an application that enhances your experiences in a museum. This application integrates eye tracking technologies and object recognition technologies to detect user's gaze on specific exhibits which can be observed when the user watches these exhibits with interests.

In our scenario, a visitor to a museum wears a head mounted eye tracker whilst strolling through the exhibition (see Figure 1). As soon as gaze on a specific exhibit is detected, the application plays an audio file that provides additional information to the user about the exhibit.

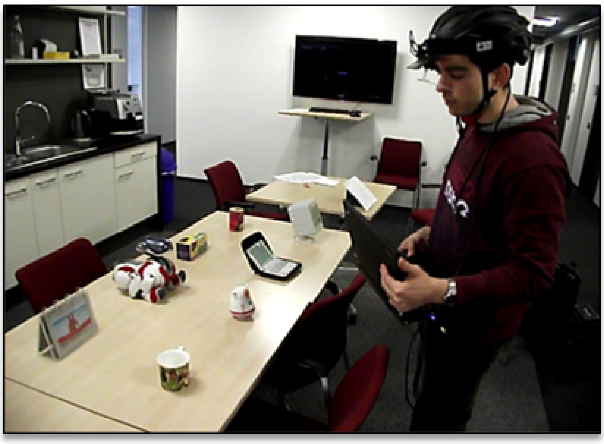


Fig 1: Sample scenario of Museum Guide 2.0

MG2.0 communicates with the eye tracker system called “iViewX^{TM} HED” [SMI] using a lean communication protocol called “UDP”. A simplyfied model of this application is shown in Figure 2. Images from the scene camera and eye tracking data are sent from eye tracking software iViewX^{TM} to the MG2.0 application and our built in object recogtion algorithm then judges whether the user's gaze exists on a specific object. If it is the case, MG2.0 starts to play prerecorded data of the specific exhibit.

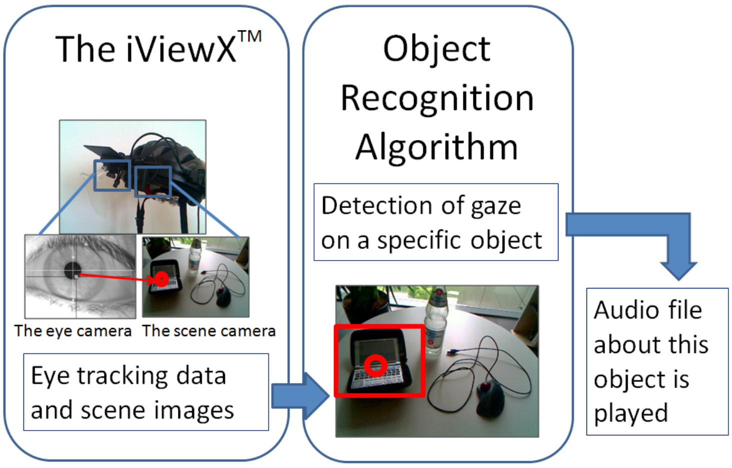
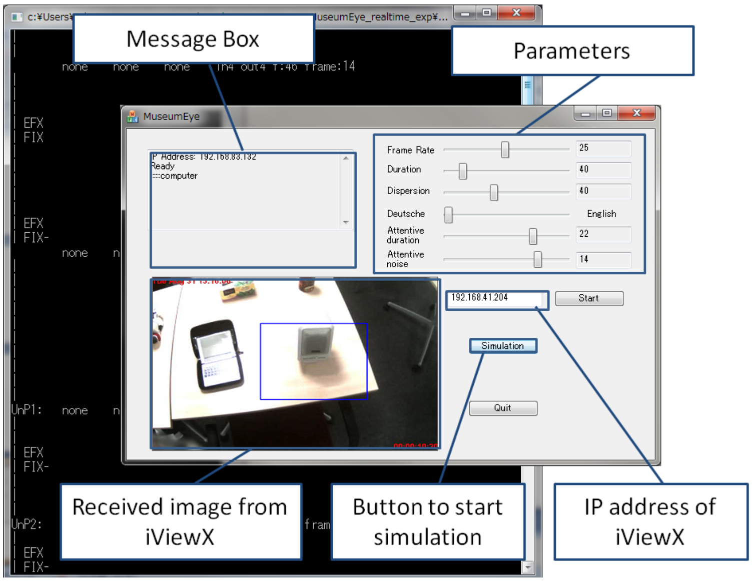

The iViewX™
The eye camera
The scene camera
Eye tracking data and scene images
Object Recognition Algorithm
Detection of gaze on a specific object
Audio file about this object is played

Fig 2: Brief model of Museum Guide 2.0

2 Museum Guide 2.0 Runtime System

The human eye is characterized by very frequent jumps from one point to another and only short timespans for which the eyes rest on a specific point. This resting of the eyes is what we refer to as *fixations* and the fast movements from one fixation to another are called *saccades*. The main purpose of the eye tracking software is to detect fixations and to synchronize the viewing direction with the images that are recorded from the scene camera. Therefore the eye

tracking software needs to be calibrated for each new user first.
Then, the runtime application of MG2.0 must be started. After succesfully loading the database file, the system prompts “Ready” in its *MessageBox* (see Figure 3).
The eye tracking software and MG2.0 communicate via IP. After starting the application (“Start” button), MG2.0 begins to receive image- and fixation data from the eye tracker.
The runtime application continuously shows the images of the scene camera and indicates the user’s gaze direction with a blue rectanglar box (its centroid is the fixation point). The cropped image of that area is then piped to the SIFT feature based object recognition framework [Lowe 2004]. SIFT (Scale-invariant feature transform) is an algorithm in computer vision to detect and describe local features in images and is well suited for content based indexing and retrieval of images.
If a known object is recognized, this subsystem returns its the symbolic name that the operator has assigned to identify this object – also refered to as “label”. However, this label does not directly reflect the object of interest, as many fixations are unconscious and may be considered as noise in our application context.
The image recognition results will rather be used to label the frames of the video stream with the label of the recognized object. All frames that belong to the same fixation will get the same label. Some frames however, will not be labeled: This is the case, if no fixation was registered a that point of time or if the confidence of the retrieval system for a known object was too low. As MG2.0 runs in a realtime scenario, some fixations might furthermore not have been analyzed as the average processing time to analyze one image exceeds the average time between two fixations.



```
P Address: 192.168.83.132
Ready
::::computer
```

Frame Rate 25
Duration 40
Dispersion 40
Deutsche English
Attentive duration 22
Attentive noise 14

192.168.41.204

Message Box
Parameters
Received image from iViewX
Button to start simulation
IP address of iViewX
MuseumEye
Start
Simulation
Quit
none none none In4 out4 T:46 Frame:14
EFX
FIX
UnP1:
UnP2:

Fig 3: Screenshot of Museum Guide 2.0 monitor application

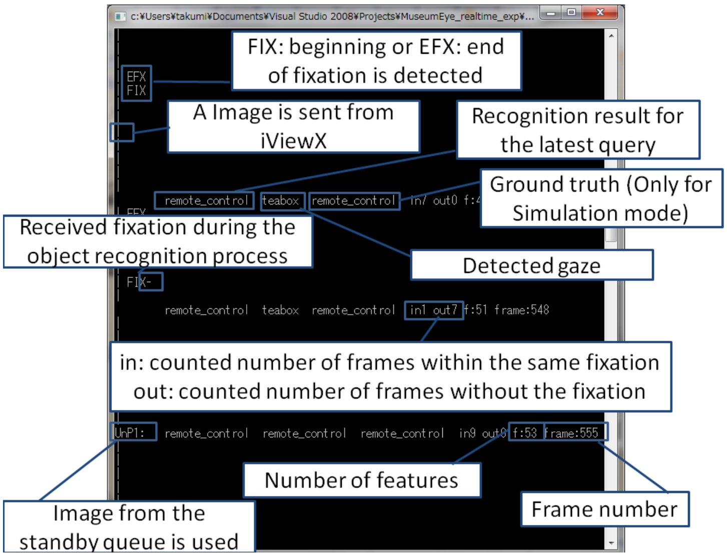
The sequence of labels of the frames is then analyzed to distinguish noise from non noise and to recognize the event of consious gaze (the trigger for MG2.0). The event of such a trigger is also monitored in the application as the name of the object is displayed in the MessageBox. For details about these analysis steps please refer to [Toyama2011]. As soon as the application detects gaze on a particular object, it plays the prerecorded voice data with information about that specific object to the user. Gaze detection is mainly driven by two system parameters which we tuned in experiment to best suit all test persons. With these parameter values, the AR starts with a delay of one or two seconds.
The operator interface also allows the modification of a set of system parameters. The individual values are preset to defaults that have been determined by a series of experiments

that we conducted during the development of MG2.0. Most mentionable here are the values for *Attentive duration* and *Attentive noise*. Attentive duration specifies for how long an object has to be watched before the users interest is anticipated and audio data is displayed.
Attentive noise specifies the maximum duration for which the user may fixate another object so that this can be considered as noise.

If the value of *attentive duration* (number of frames) is chosen to small, MG2.0 will likely present audio data for an object that the user does not watch consciuosly. If it is to large, the presentation starts with longer delay or not at all. The suitable value has been evaluated in experiments with different users and the value *22* is set as the default value.

Similarly, if *attentive noise* if chosen to small, the system resets the counter too often, that has to reach the *attentive duration* value. It might thus happen, that the system does not recognize gaze at all or only if the user acitvely concentrates to one object in an unnatural way. If that parameter value is chosen to large, even long fixations to one object are treated as noise and the system reacts more lazy. The experimentally evaluated default value is 18.

Whereas the user interface of MG2.0 presents audio data when a gaze event is recognized, the console window of the application presents much more information that is helpful for the developers and operators of the system to analyze and debug any kind of malfunction and to tune the system to achieve best performance. Figure 4 shows a screenshot of the console with annotations to the kind of information displayed.


c:¥Users¥takumi¥Documents¥Visual Studio 2008¥Projects¥MuseumEye_realtime_exp¥...

```
EFX
FIX

      remote_control  teabox  remote_control  in7  out0  f:4
EFX

FIX-
      remote_control  teabox  remote_control  in1  out7  f:51  frame:548

UnP1:  remote_control  remote_control  remote_control  in9  out0  f:53  frame:555
```

FIX: beginning or EFX: end of fixation is detected

A Image is sent from iViewX

Recognition result for the latest query

Ground truth (Only for Simulation mode)

Received fixation during the object recognition process

Detected gaze

in: counted number of frames within the same fixation
out: counted number of frames without the fixation

Number of features

Frame number

Image from the standby queue is used

Fig 4: Screenshot of the console of MG2.0

3 Populating the Database with Exhibits

To demonstrate MG2.0 under most realistic conditions, one also needs a real world environment that at least simulates a museum with its exhibits. With our intended application in mind and the requirement to recognize objects based on their visual appearance, we had to find decisions about the type of exhibits that we like to recognize in two dimensions: 2D- versus 3D-objects and small versus large objects.

2D-objects like paintings or photographs would certainly be the easier choice. However, if our approach manages to deal with 3D-objects (e.g. sculptures), we can right away recognize 2D objects as well. 3D is more challenging, as the user looks at these objects from many different perspectives and the captured camera image of the object, as taken by our camera, looks much different from each direction.

For our application we decided to face the more challenging of 3D-objects. To overcome the mentioned variety of object appearance, pictures of each object need to be taken from many perspectives. All images of the same object need to be indexed with the same unique label for that exhibit.

Slightly similar to this are the implications of the decision for either small or large objects (no matter whether 2D or 3D). When facing a large object, the visitor will typically stand close and his eyes scan the objects different regions step by step. When we like to index such a large object, we need to point the camera to the many different areas and label all images just as we did for the different perspectives.

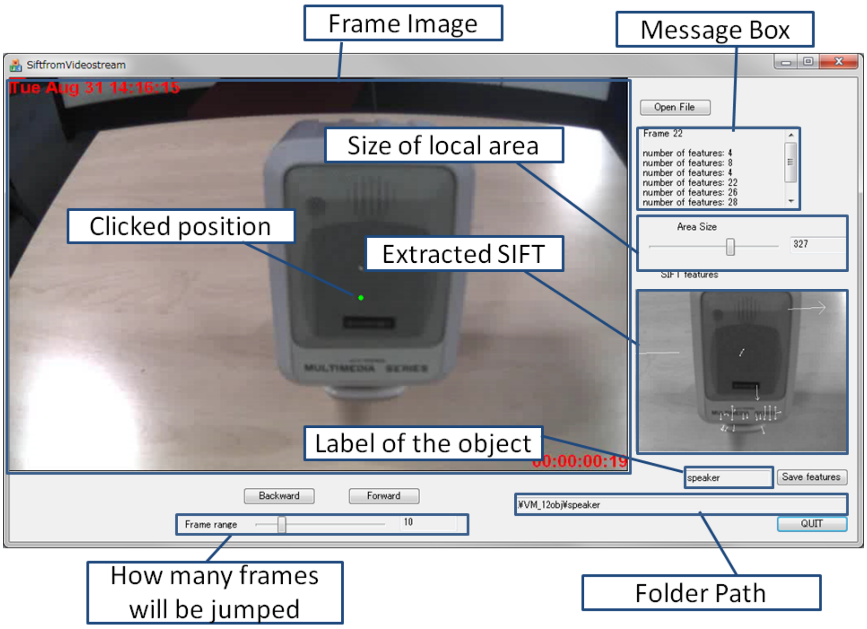
In our simulated museum scenario however, we limited the sample exhibits to small 3D objects (see Figure 1). Ideally, these objects in the museum should be arranged in a way that neighboring objects are not too close to each other so that objects do not overlay in the camera image for most perspectives.

For each object in our exhibition we perform the following steps:

1. Record a video file of the object that should be added to the database. We use the camera of the eye tracker to obtain the most similar images (wrt. e.g. brightness, contrast, resolution) as during runtime.
2. Place the object to its intended place in the museum
3. A person wearing the eye tracker walks around the table, thereby directing the scene camera to the object (Eye tracking is not required here).
4. Record the video data with the iViewX™ software.
5. Save the video with a filename matching the label of the object (e.g. “speaker”)

If the exhibits are exposed to natural daylight (not purely artificial light), it is recommended to also record the video under different lights to extract sufficient variation of SIFT features for recognition.

After that step we have one video for each exhibit in our museum. To populate our database with the SIFT features of the art objects, we have to process each video with MG2.0 Indexing Tool. A screenshot of this application is shown in Figure 5.



```
Frame 22

number of features: 4
number of features: 8
number of features: 4
number of features: 22
number of features: 26
number of features: 28
```

Area Size: 327

Frame range: 10

speaker

¥VM_12obj¥speaker

Fig 5: Screenshot of the MG2.0 Indexing Tool

With this tool, the operator can navigate through the video (in different speed, frame by frame as well as forward and backward) to search and select specific still images that appear to be ideal for the indexing. Rather than using the eye tracking information as done at runtime, during the indexing phase the operator has to manually select a point in the still image. The application now extracts SIFT features from the local area whose center is the clicked position. The operator can define the size of the local area with a slider. The extracted features within the image are shown as white arrows in the smaller image box at the right. These features can then be saved to the database. The database itself is represented by a defined path in the filesystem, following certain naming conventions. Just like the label of the exhibit defined the basename of the video, this name is also reflected in the database as it contains subdirectories with that name. All features for this object as well as the still images are stored in that subdirectory. To achieve a good performance, the total number of features should be even among objects. As a rule of thumb, the largest number of features from one object should be double to the smallest one. Objects that have lesser visual structure and thus fewer features should therefore be indexed with more images.

The last step to prepare our system for a specific museum is the definition of the audio files that are played to the user if gaze to an object is detected. The recordings can be done by an external tool (e.g. voice recorder). These files simply need to follow the same naming convention for the basename as the video file, but with the extension “wav”. That file then has to be moved to a folder which contains all the audio files. Different such audio folders can be defined for multiple languages.

4 Evaluation – User Study

To evaluate the usability of the complete system, we conducted a user study with 23 users. The users were recruited from the staff of the DFKI knowledge management department, and most of them were students in the age of their mid 20-ies. They are familiar with ICT technology and have experienced the established AR systems in real museums already, but they have no specific affinity to museums. As the object recognition and anticipation of user interest based on the eye-gazing information have been the central aspect of our research, no real museum was involved in the studies. Instead, we rather built “our own museum” out of daily life objects.

The users were asked to stroll in our museum with two different guide system. One is our Museum Guide 2.0 and the other is an audio player based guide system. Audio player based museum guides are currently used in most of the museums and therefore provide a good basis of comparison with existing technology. Usually exhibits have a tag number in front of them and the users have to select the corresponding audio track from the audio guide to get more information about that exhibit. The same setup was used in our experiment by assigning a tag to each of the twelve objects in our museum and storing the corresponding audio information with the same tag in the audio player. The users were asked to freely move in the museum and get information about the object they are interested in with the help of the audio player. After the users finished their round with the audio guide, they were introduced to the eye tracker and the eye tracker was configured and calibrated for each user. This whole setup process for each user took 10 seconds in the best case, but also 5 minutes in the worst case (e.g. if a user had a different dominant eye than his predecessor and the eye camera thus had to be mounted to the other side or if the pupil was not recognized immediately). If the calibration was not done properly, the system could not detect the accurate gaze position in the scene, which then causes wrong trigger events for MG2.0.

When this setup step was done, the users were asked to take another round in the museum wearing the eye tracker. Whenever the users gazed at an exhibit and gaze on exhibits was detected, Museum Guide 2.0 played a pre-recorded audio file to provide more information

about the gazed exhibit.
When the users finished their round with Museum Guide 2.0, they were given a questionnaire to assess different aspects of the system. A summary of user responses to the questions comparing the gaze based interface with the traditional audio player interface is shown in Figures 6-10. Since the eye tracker used in the study has several hardware constraints (such as uncomfortable helmet, chin rest, etc.), we referred only to a “gaze based interface (device)” in the questionnaire to judge the real potential of gaze based information provision. The results show that most of the users would prefer to use a gaze based device as compared to an audio player when they go to a museum. Another interesting result was that although many users were satisfied with the traditional audio player, the mean opinion score (MOS) for Museum Guide 2.0 was 4.3 as compared to 3.2 for an audio player.

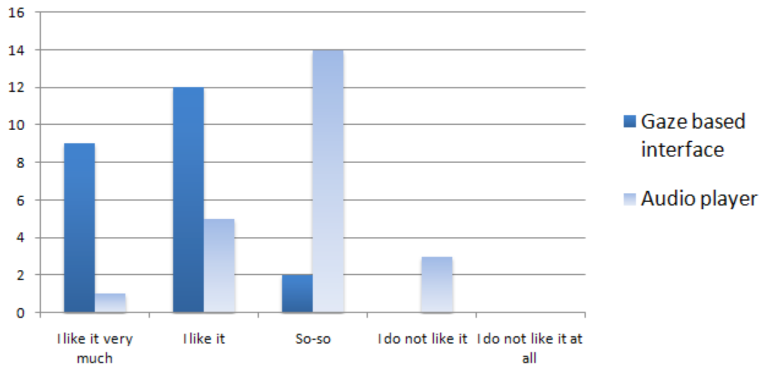


Fig 6: Responces in the user study for the question: How much do you like a gaze based interface (or a traditional audio player) for getting information?

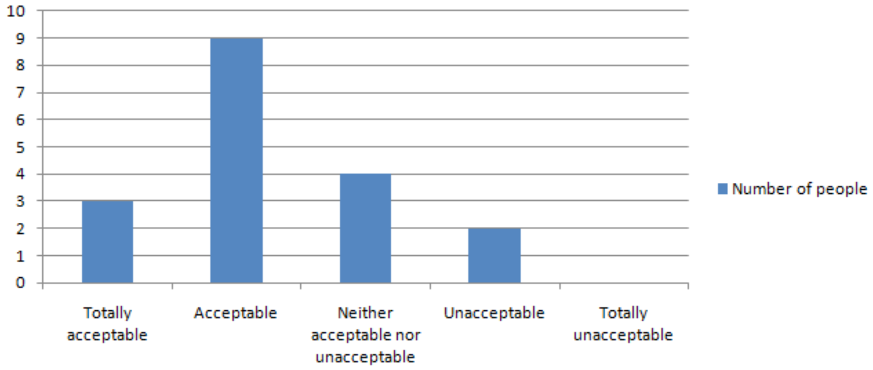


Fig 7: Responces in the user study for the question: How much was the calibration process acceptable for you?

We must admit, that the users of our study have to some degree been biased in a positive way: The study has been conducted by the developer himself and the users have been colleagues or even friends. Also, we argued that the users shall abstract from the need to wear the helmet with the built-in eye-tracking device. But we did at least not register negative tendencies which would have been taken as a knockout for the idea of MG2.0. Nonetheless, the questions reflected in Figure 8 and 9 were of higher importance to us, because the more objective answers reflect the central aspects of our research: the anticipation of the users interest to some specific object based on gaze information.

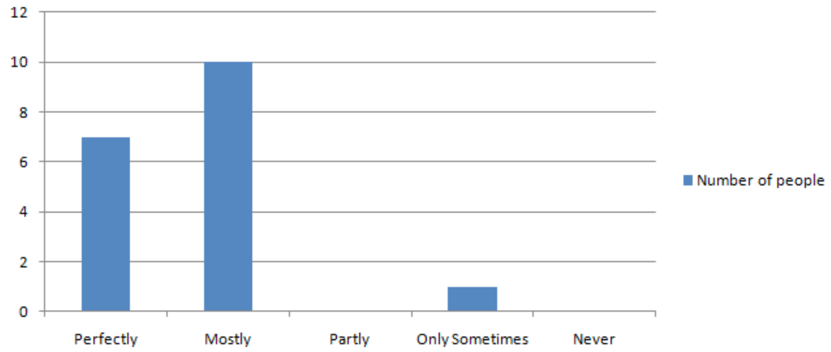


Fig 8: Responces in the user study for the question: Did you get the information against the object you want to know?

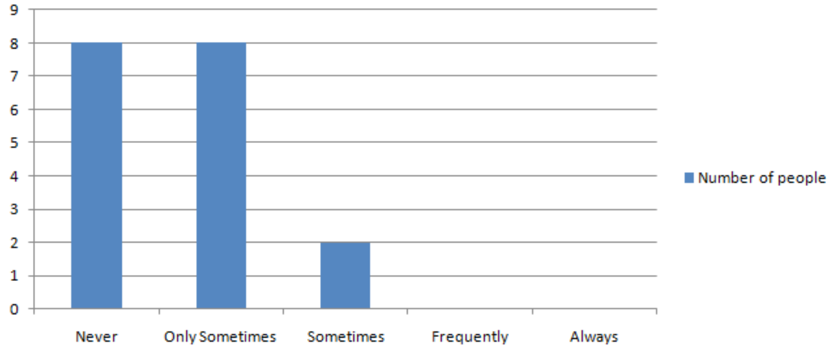


Fig 9: Responces in the user study for the question: How often did you get the information against the object which you were NOT interested in?

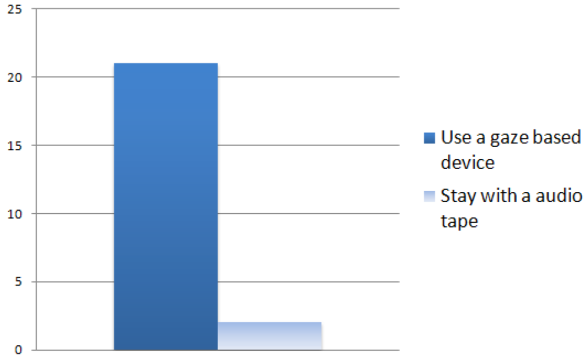


Fig 10: Responces in the user study for the question: What would you like to use when you go to a museum (Ignoring the hardware constraints)?

5 Conclusion and Outlook

This paper presents a novel approach to provide Augmented Reality to users based on gaze information. The technologies involved are eye tracking devices, image based object recognition and own approach towards anticipation of the users' interest. Our aim was to demonstrate the feasibility of such technology in practical applications (without considering location information – similar to Google Goggles – which might further boost recognition performance). And we selected the domain of museums, specifically inside exhibits of static objects, as these environments provide the neccesary settings to achive best object recognition results. However, we are aware, that a system like MG2.0 would increase the costs for setup and maintenance in comparison to state of the art AR technologies. At the other hand, MG2.0 might specifically serve elderly people or people with motorical disabilities, as they do not require any explicit manual interaction. MG2.0 thereby resembles a personal human guide. Pros and Cons of a system like MG2.0 thus need to be balanced carefully.

Today, eye-tracking devices are still rather expensive and MG2.0 is hosted on a laptop PC that needs to be carried along. But technology evolves quickly. With an increased demand for those devices – as it might be raised by applications like MG2.0 – larger production series will lead to reduces prices and the hosting computers will soon have the size of a smartphone. In parallel, the eye trackers become smaller as demonstrated by SMIs new Mii device. Rather than being mounted to a helmet, this device integrates its components (including scene camera and eye-tracking cameras for both eyes) in the frame of almost ordinary glasses. As announced by SMI, that new device also reduces the mandatory calibration step from today five points to just one point at which the user has to look during this step.

Acknowledgements

This work was partially funded by the BMBF (German Federal Ministry of Education and Research), project Perspecting (01 IW 08002). We also wish to thank SensoMotoric Instruments (SMI) GmbH for providing us an iView X^{TM} HED mobile eye tracker for building the application and conducting all experiments.

References

Takumi Toyama, “Gaze Guided Object Recognition Using a Head-Mounted Eye Tracker”, *Master Thesis, Graduate School of Engineering, Osaka Prefecture University*, March 2011

Lowe, D. G., “Distinctive image features from scale-invariant keypoints”, *International Journal of Computer Vision* 60, 2, 91–110, 2004.

Santos, Pedro and Stork, André and Linaza, Maria and Machui, Oliver and McIntyre, Don and Jorge, Elisabeth. “CINeSPACE: Interactive Access to Cultural Heritage While On-The-Move”, Online Communities and Social Computing, *LNCS Volume 4564*, pages 435-444, 2007

Agamemnon, EU co-funded undert the IST 6th Framework Program of the European Commission, information available online: http://services.txt.it/agamemnon/

SMI - SensoMotoric Instruments, “iView X™ HED – The mobile, head-mounted eye- and gaze tracker”, http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/iview-x-hed.html