Usability of Eye Tracking as a User Research Technique in Geo-information Processing and Dissemination

GFM MSc Research

Rozita Razeghi

March 2010

Usability of Eye Tracking as a User Research Technique in Geo-information Processing and Dissemination

by

Rozita Razeghi

Thesis submitted to the International Institute for Geo-information Science and Earth Observation in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: (fill in the name of the specialisation)

Thesis Assessment Board

Chair: Prof. Dr. M.J. Kraak
External examiner: Dr. M.L. Noordzij
Supervisors: Dr. C.P.J.M. van Elzakker, and Dr. C.A. Blok
Co-supervisor: Mr. I. Delikostidis

ITC

INTERNATIONAL INSTITUTE FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION
ENSCHEDE, THE NETHERLANDS

Disclaimer

Abstract

Nowadays, different fields of science are growing fast, and the advancement of technology as a tool, accelerated this speed. Technological advances in the geo-domains, like other disciplines had a great role in development of geo-information science, and every day newer systems and techniques of geo-spatial data processing and dissemination are assessed. One of newly considered and essential aspects in evaluating the technology is the usability of the products, as well as methods and techniques for doing use and user research in different disciplines including geo-information processing and dissemination, and eye tracking (ET) is one of this techniques. It records the focus points of a person while they see e.g. an interface. The main potential of ET in preference to other user techniques is to reveal hidden cognitive processes of human behaviours; however, there are some essential objections to ET, e.g. it cannot say anything about why users are looking at some object. Taking into consideration the cost-benefit perspective, the usability of the technique as a whole, needs more investigation which is the main objective for this thesis.

For this purpose and in order to get a general overview regarding the usability of ET with different systems, we designed 2 case studies with 2 main types of ET systems (mobile and stationed). Also, in order to assess the effects of the thinking aloud technique on ET (which can compensate for the objection of ‘why users are looking at some object’) and to find out the most resultful combinations of these techniques, we applied 2 homogeneous user groups with different methods of usability in each case study. These combinations were the use of ET alone and ET with thinking aloud (together with other user techniques).

The first case study, assess the usability of ET in providing a methodology for selecting salient landmarks in the analyze requirements phase of a pedestrian navigation system. Using different methods, users were asked to navigate on a predefined path and find a destination, while their gaze data were recorded. Likewise, the second case study deals with evaluating the usability of ET in providing a methodology in the produce designs solutions phase of a geo-web application which is a prototype for visualizing iceberg data. Depending on their groups, users were asked to find the answer to some questions regarding the icebergs using the application, and to apply TA or not. During the design, data collection and analysis of these 2 case studies, some issues related to calibration, data clustering, analysis and interpretation were discovered and discussed. Also, the 2 applied methodologies in each case study were discussed and compared against each other regarding the main variables of usability testing which are efficiency, effectiveness, and satisfaction.

The results of the first case study showed that ET indeed can provides information about ‘human cognitive aspects’ in selecting salient Landmarks. Also, using the combination of ET and TA proved to be more informative for this application. Unfortunately, due to some technical problems, the data of the second case study were not analysed properly to provide certain results; however, the combination of ET and TA was recognized to be more informative for this case study.

Acknowledgements

All the worship and adoration to almighty God, the beneficent, the merciful as the first and everlasting teacher, leader and friend in my entire life; without his divine grace, the successful accomplishment of this research was impossible.

My sincere thanks to the EU Erasmus Mundus consortium for awarding this scholarship to me to pursue such an exceptional graduate program. I would like to convey my acknowledgment to all my teachers and to the coordinators of this programme in ITC.

My utmost gratitude goes to my supervisors Ms. Dr. Connie A. Blok and Mr. Dr. Corné P. J. M. van Elzakker for their precious inputs, critical advises and expert guidance. All your great ideas helped me maintain the focus during all phases of the research for which I thank you.

I am grateful to my technical advisor, Mr. Ioannis Delikostidis for sharing his knowledge, invaluable suggestions and his technical support during this research.

I like to extend my appreciation to GFM program director and course coordinator,
Mr. Gerrit C. Huurneman and Ms. Dr. Ir. Wietske Bijker for their kindness and moral support towards me during my stay in the Netherlands.

I especially thank GFM-2008 classmates who voluntarily participated in my case studies during this research. You were wonderful classmates indeed. Eighteen months of living and studying together yielded comradeship, fellow-feeling cordiality and unity.

I am forever grateful to my beloved parents, Hassan Razeghi and Mehri Jabbari, for their continuous inspiration in my life, and for guiding me in a way to always follow moral principles. Special thanks to my beloved uncle Dr. Hamid Jabbari who never let me feel alone or unsupported here in Enschede.

My heartfelt thanks go to all my loving family members for their warm affection, support and continuous encouragement. You are the best!

Table of contents

List of figures

List of tables

1. Introduction

1.1. Overview

Today, different fields of science are growing fast, and the advancement of technology accelerated the growth of science in various domains. Like in other fields, technological advances in geo-informatics have played a great role in the development of science in the geo-domains, and every day new solutions for methods, systems and techniques of geo-spatial data processing and dissemination are being assessed. Some of the aspects of growing importance in the assessment of the technology which has been considered in geo-informatics and other disciplines are the user and his needs, demands, and limitations in utilizing the technology. For a long period, there was limited attention towards the user as part of investigating the technology and it was not clear whether the created products in geo-informatics were usable or made sense. Today, however, there is gradually more attention towards testing the usability of the products developed for geo-information processing and dissemination as well as more attention for methods and techniques for doing use and user research. These methods are used to evaluate products by testing them on users and measure the product's capacity to meet its intended purpose. One of the techniques which has reappeared recently, although it was already used in several domains in the past, is eye tracking (ET), which tracks the eye movements of a person as they see (for example) an interface, and records the focus points at which they looked. Typically, an ET system works by reflecting infrared (IR) light onto an eye, recording the reflection pattern with a sensor system, and then calculating the point of gaze using a geometrical model. Once the point of gaze is determined, it can be visualized and shown on a computer monitor as a moving cursor.

The origins of ET are over a century old; it started as a very invasive technique and gradually turned into an almost completely non-invasive one. ET has various applications in different domains like visual system analysis, psychology, cognitive linguistics, product design, human-computer interaction (HCI) and many others. ET application in HCI could be divided into usability assessment and a newer application which uses ET as a direct control medium within a human-computer dialogue. In usability measurement, the recorded gaze data during system use typically would be analyzed offline. Using gaze data as one of the several inputs to a system, on the other hand, includes finding proper ways to react judiciously to eye signals in real time and avoid over responding; the technique is applied for hand-busy or disabled users.

Recently, ET has been applied to some extent in usability research in order to improve the design of different applications in geo-information processing. Researchers have reported that ET enhances usability testing by combining it with other conventional techniques such as interviews and questionnaires (Çöltekin et al., 2009). They also declared that ET reveals hidden cognitive processes, which are valuable sources of information about human behaviours or usability problems that may not be gained by conventional usability data collection techniques. ET is also considered to be useful for testing hypotheses about the design (Cooke, 2005).

ET is mostly applied in the last stage or design evaluation of a User Centered Design process in the geo-information domain, but it may also be applied in the first stage or requirement analysis stage of the UCD process.

1.2. Motivation and problem definition

As a fairly noticeable, real time measure of visual and cognitive information processing behaviour, ET was believed to be a promising technique over the past 50 years. Still there are some uncertainties about its usability regarding its availability, intrusiveness, robustness, costly performance and analysis with no warrant of getting extra knowledge or usability of its resulting data; these kinds of problems have held it up at such a merely promising stage. For this purpose, some issues related to clustering, analyzing and interpreting the data, which partly caused this slow start, are discussed below.

ET is the process of measuring either the point of gaze or the motion of an eye relative to the head in order to determine eye movement patterns of a person. An *eye tracker* is a device for measuring eye movement which comes in two main types: *head mounted* ET systems for portable applications and *stationed* ET systems for document based and computer screen studies. An eye tracker typically records eye location, gaze location (usually as a 2D point), the time and the duration for each sample. Depending on the sampling rate, and the duration of the session, the amount of data may become huge and complex. *Clustering* is a solution to reduce this high data volume by classifying the scattered data points into somewhat more regular categories. In data analysis, the first step normally is to cluster data. Clustering ET raw data mostly includes differentiating between *fixations* (pauses over informative regions of interest) and *saccades* (rapid movements between fixations). This can be realized by separating raw fixation points and collapsing them into a single representative tuple. By separating fixations, raw saccades would be separated implicitly. Most data processing of human vision occurs during fixations, while during saccades almost no information absorbed by the eye. Hence, saccades are not often included in visual processing. Furthermore, there are some unwanted eye movements during a fixation, and extracting fixations reduces this data noise resulting from smaller eye movements (Salvucci and Goldberg, 2000).

A wide range of methods and techniques is applied for eye detection, data clustering and analysis at different stages of conducting the ET technique. Choosing the best algorithm to use for a specific task is a challenge. While different algorithms can be used to perform the same task, each algorithm may produce a different result. Furthermore, the same algorithm may produce different results by changing its parameters, which also leads to different interpretations of the same ET data. Due to the development of technology, there are some available software applications which provide analysis tools and automatic methods for extracting the fixations and saccades from data. However, because of the strong dependency on the choice of their parameters and also a lack of transparency regarding the algorithms (Lankford, 2000), sometimes there is a need for more customized routines to control the data.

Conducting *analysis* of the clustered data properly, as the next step, requires selecting relevant metrics as well as appropriate operations to evaluate them against each other or other techniques. These operations could be done by the present ET software or developing personal tools based on requirements.

Depending on application, a variety of different eye movement measurements are feasible. Common analysis metrics of ET data include fixation, saccadic velocities, saccadic amplitudes, and various transition-based parameters between fixations. However the major part of ETanalysis is related to separating and labelling fixations and saccades. Interpreting processed data as the last stage of an ET study is not always an easy job, and it is to some extend based on the opinions and understanding of evaluators (Ehmke and Wilson, 2007). For instance existence of the long fixations over some particular part of the stimulus can be interpreted either as user’s interest or confusion at that particular part. ET results are not clear enough to be interpreted in isolation. In practice, ET is often combined with other techniques of user research to assist the interpretation process. However, interpreting integrated ET data with data from conventional techniques can be even more challenging (Jacob and Karn, 2003). For instance, the support of data of some conventional user research methods like thinking aloud protocols (TA) resulting from users who verbalize their thoughts during task completion makes the interpretation

process to some extend subjective. In addition, while ET can enhance usability testing by combining it with other traditional techniques, still there are some disagreements on how they can be joined to achieve optimum results. Looking at the sensitive nature of ET experiment proves that concurrent use of TA can cause difficulties to the users as they are verbalizing ongoing cognitive processes that may be subconscious (Ball et al., 2006). However, some researchers e.g. (Brodersen et al., 2002) proved that concurrent use of TA and ET could lead to valuable results. There is also a problem regarding the effect of the interface context on the user and a limitation in deductable conclusions about cognitive processing which are obtainable from ET data. For instance, 'a given cognitive event might reliably lead to a particular fixation, but the fixation itself does not uniquely specify the cognitive event'
(Hayhoe and Ballard, 2005).

The ET technique provides a different approach in usability research, not as a replacement, but to supplement conventional usability testing. It reveals behaviours that would be difficult to obtain through other test measures. Taking into consideration its potential benefits versus the vital range of problems like cost, calibration, handling data, complexity and combination of methods, the usability of the technique as a whole still needs more investigation.

1.3. Research objectives

The main objectives of this research are to investigate the usability of ET and the resulting data in different stages of User Centered Design of a product or tool in geo-information processing and dissemination, and to present guidelines for a useful implementation of the ET technique in use and user research in the geo-domain. These main objectives can be divided into some categories of sub-objectives:

- To present information regarding the general structure of the ET research technique.
- To propose proper techniques for integrating data from applying the ET technique with other techniques of user research.
- To propose proper (visual/automatic) techniques of data clustering, analysis and interpretation of the processed data in order to improve the usability of different geo-applications or geo-information handling.

1.4. Research questions

Questions are organized into four groups: first general questions for technique identification, and then more specific questions like data collection questions, data clustering and analysis questions, and data interpretation questions.

- General questions

    - What is ET? What is it applied for? What are possible applications and capabilities in different domains? What are the related experiences in the geo-domain, and what were the comments on the usability of the resulted data?
    - What are the advantages, disadvantages and current pitfalls of the technique? Are the disadvantages manageable?
    - In which stages of the UCD process of the geo-information domain can the technique be applied?
    - What other factors can affect (improve or bias) the results? Does a learning effect, environmental condition, test duration or other factor modify the test results?

- Data collection questions

    - What kinds of raw data are collected by an ET system?

- How is the ET system applied? What is a typical test procedure? What does the test person/experimenter typically do during the test? Does the device type (mobile/fix) effect the execution procedure?
- What kind of different components do ET systems include? Do the equipments differ for different systems (mobile/fix)? What do all components do?
- What are the different capabilities of the two (mobile/fix) systems? Can mobile systems be applied for document based and computer screen studies as well?
- What is the general workflow of ET systems in order to record the track of the eyes?
- Should ET be combined with other usability techniques? Which method (alone/combined) is preferred as a user research technique? How should ET be combined with other usability techniques?
- What is the use and the effectiveness of using the ET technique compared to other comparable user techniques like audio-video recording which they cannot provide? And which method is preferred?
- Which preparatory activities, assistance, further equipments or other material are required? What are the costs involved (in terms of manpower, equipment, time)?

• Data clustering and analysis questions

- What are the impediments of clustering data?
- What (visual/automatic) clustering and analysis tools are currently available? Can available softwares meet the needs of clustering and analyzing data in the geo-application domain? Do the softwares differ regarding the system types (mobile/fix)?
- What are the defined ET metrics for analyzing the data? Which metrics (fixation, saccade, AOI and etc) are proper to be analyzed in different geo-applications?
- What other measures (video, sound, mouse click, etc) can be combined with ET metrics to improve analyzing the data via available software? Can we use combinations of these measures via available software?

• Data interpretation questions

- What kind of results can be obtained with this technique in order to improve the usability of geo-information applications?

1.5. Thesis structure

Chapter two has been devoted to the review of exiting information about characteristics of the ET technique including definitions, usage in different stages of UCD, application in different domains as well as in the geo-domain, advantages, limitations and pitfalls of the technique. It also describes the different stages of applying the ET technique including data collection, clustering, analysis and interpretation. Knowing the present potential possibilities, capabilities and drawbacks of the technique provided in this chapter, along with considering available facilities in ITC, guides us to the design, implementation and data analysis of two case studies in order to assess the usability of ET for a mobile and stationed system respectively, in the next two chapters (chapter 3 and 4). The first case study describes a new application of the ET technique in the geo-domain, which involves the requirement analysis phase of the UCD of a pedestrian navigation prototype for selecting salient landmarks. In order to get a more general overview of the usability of ET in the geo-domains, another case study is planned and discussed with a stationed system in the same chapters. It assesses the design solution of a geo-web application for visualizing the ice berg data. Finally, based on these case two studies, in chapter five, we have come up with some conclusions and recommendations for future researches and applications of the ET technique in the geo-domain.

2. Literature review

2.1. Overview

This chapter is devoted to literature review of general characteristics of the ET technique. It starts with describing general concepts regarding usability testing and its different methods and techniques including ET. Then in section 2.3, ET and a brief history are defined more specifically. Section 2.4 and 2.5 shortly describe some applications of ET in different domains and also in geo-information processing and dissemination. The next section (2.6) is devoted to review and assessment of various existing methods applied in different stages of a typical ET study. An ET study usually consists of different stages including data collection, data clustering, data analysis and data interpretation. Depending on the applied hardware and software or processing methods, there are different alternatives for implementing each stage. For instance, data could be collected via different device types. Likewise, the collected data could be clustered by different methods etc. Section 2.6 defines and describes each of these stages and the alternative methods for each stage in details. Finally, in section 2.7 the main potential, drawbacks and limitation of ET are identified.

As mentioned in chapter one, there are different method and techniques for usability testing, and ET is one of this techniques. The next section gives more details about definitions and classifications of these method and techniques including ET.

2.2. Techniques of usability testing

This section defines basic concepts which are related to the usability. It briefly describes User Centred Design (UCD), usability testing and its usual methods and techniques. As already mentioned, users and their demands are one of the vital factors in the evaluation of any types of technology including the products which are related to the geo-information processing and dissemination.

Usability is a gauge of usefulness of an interface for its end users. The interface is applied to the hardware, software, online help documentation or any other element that shape user experience with the product; examples in the geo-domain are like a digital/paper map, geo-website, PDA, etc. Usability is defined as ‘the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use’ (ISO 9241-11). It means the users of a product should be able to accomplish their tasks easily and quickly (Barnum and Dragga, 2001). It is also defined as the quality in use along with efficiency, reliability, maintainability, functionality and portability of software products (bevan, 1999).

Usability engineering is a basic term applied to a set of activities in order to create user centred product designs which can use either experts or non-experts as evaluators. Usability testing is a part of usability engineering which is defined as ‘the process of learning from users about a product’s usability by observing them using the product (Barnum and Dragga, 2001)’ During execution of a predefined task in usability testing, the user’s reactions, behaviours, errors, and self-reports are carefully observed and recorded by a usability engineer. Then these data are fed back to the designer to improve interface usability in (re)design. Usually it is an iterative process, meaning that after redesigning, the product would be tested again till finally the developer is satisfied that the users’ desires are met. The aim of the usability testing is to estimate the amount of effectiveness, efficiency and satisfaction of a product’s user.

Usability testing includes different methods like heuristic evaluation, cognitive walkthrough, task analysis, remote testing, etc. Each method has one or more techniques embedded into it. For instance, task analysis can make use of different techniques like interviews, questionnaire, focus group, etc.

Usability techniques can be either quantitative or qualitative (URL6). The qualitative techniques give non-numeric description. Some techniques of qualitative methods are like direct observation, introspection and TA. Quantitative techniques which give numeric or statistical output are like questionnaires, performance measurement. ET is also one of these quantitative techniques that can be embedded into different usability testing methods. Since the main objective of ET technique is to reveal cognitive behaviours of an end user of a system, it is usually applied in methods which employ a real user to carry out a task. However, in some applications it is possible to apply it for the experts as well. An example of such applications is the use of ET for the purpose of differentiating between novice and more experienced users while applying a system and then using experts' eye patterns for teaching novice users (Jacob and Karn, 2003). Usability testing methods can generally be studied under three groups of testing, inspection and inquiry (URL5). In testing approach, usually typical tasks are executed by representative users using a prototype or system, and then the results are assessed by evaluators. Commonly used testing methods are like remote testing, shadowing and TA. Inspection is based on assessment of usability-related aspects of a system by the experts (and users). Inspection based methods are like heuristic evaluation and cognitive walkthrough. Inquiry is based on both observing users using the system in a real work and communicating with them (verbally or in written form) by evaluators. Examples of Inquiry are focus group and field observation. Table 2.1 and 2.2 include the descriptions of some important usability testing methods and techniques.

Table 2.1: Some important methods of usability testing.

Usability testing method	Definition
Heuristic evaluation	Experts evaluate a product according to some accepted principles
Cognitive walkthrough	Experts walkthrough the steps and actions required to accomplish a task and compile potential issues
Pluralistic walkthrough	Experts walkthrough the actions required to accomplish a task as a group asking themselves a set of questions
Task analysis	Is the process of learning about the product by observing users in action

Table 2.2: Some important techniques of usability testing.

Usability testing technique	Definition
Interviews	To enquiry user verbally after completing a task
Questionnaires	To ask user to write down their ideas in answering questions about the test
Thinking aloud (TA)	To ask user to say aloud his thoughts while completing a task
Retrospective think aloud	To ask user about what he thought after completing a task
Direct observation	To observe user's behaviours while completing a task
Video-recorded observation	To video record user's behaviours while completing a task
Screen-logging observation	To record user's interactions with interface via data logging software applications
Focus group	To record a moderated discussion among potential users of a product
Eye tracking	To capture user's point-of-gaze while completing a task

Although usability testing can be conducted through several methods, for developing a more usable system it is recommended to use a UCD approach (figure 2.1). UCD involves a set of methods which employs a user's interaction for design development as well as for the final production through an iterative process. UCD approach supports the entire development process cycle of user centred activities in order to create more usable product and with less cost. The three stages of the UCD includes *analysis requirement*, produce design solutions and *evaluate designs*. It starts with the analysis requirement phase, which deeply investigates users' requirements and encountered problems with existing products and with considering context of use and task. It is also possible to carry out this stage without a pre-existing product and only by a careful assessment of the requirements that a system is to fulfil as well as exploring likely problems that user might encounter while using the product. The results of the assessments in this stage leads to development of the first prototype in the next stage which is produce design solution phase. This developed prototype then would be tested again with representative users in evaluation designs phase, with the aim of meeting users' requirements, and again it is back to the product design solution phase (for redesign) in an iterative way. Different usability testing methods can fit into different stages of the UCD process. For instance, task analysis can take place only in the first stage of the UCD process. Likewise, heuristic evaluation can take place both in the first and the last stages of the UCD process (URL6).

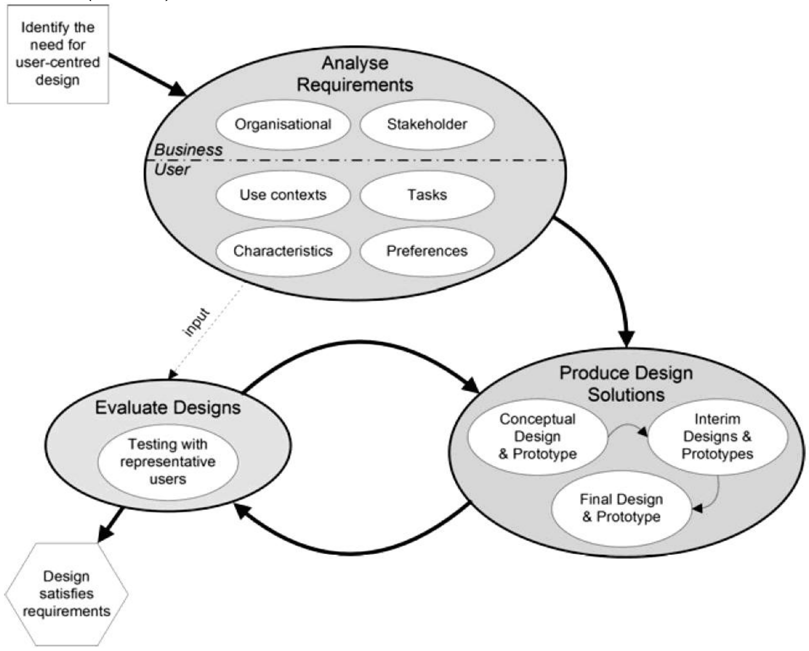


Figure 2.1: A general diagram of the User Centred Design process (Van Elzakker and Wealands, 2007).

Depending on which stage of development life cycle the product is fulfilled, different types of usability testing is conducted. In order to examine preliminary designs as models or prototypes, exploratory tests are performed. As a midway into the development cycle, assessment tests are carried out to assess whether the design concept was implemented. Finally, validation tests are carried out to confirm whether the final product is comparable to some standard or other products. Compared to earlier stages, the later the test is conducted, the less qualitative and more quantitative it would be (Bojko and Schumacher, 2008). As in comes to ET, there are few cases of applying this technique in the first stage of a product development cycle in the geo-domains (Davies and Peebles, 2007) and it is often utilized while the target product is in the prototype stage. Prototypes are tested against each other and competitors to examine for instance which specific elements are associated with high visibility and appeal (Brodersen et al., 2002). ET has a high potential in giving insight into cognitive aspects which otherwise could hardly be gained. This potential is an important factor in evaluating a product in both the analysis requirement stage and the evaluate designs stage which may lead to discovering extra information about the requirements of the user.

2.3. ET user research technique

ET is an objective technique of measuring an individual's eye movements, including where their eyes are looking at any given time and the sequence in which they are shifting from one location to another by means of a device called eye tracker. This technique opens a new opportunity to look at humans motivation from another dimension added to the already mentioned research techniques. It helps to reveal hidden cognitive processes of human behaviours which cannot be obtained by other techniques; however, there are several objections to this technique (see section 2.7). For instance, ET does not say anything about why users are looking at a particular point.

Eye movement's investigation precedes the wide-spreading deployment of computers by one century, and it gradually evolved from very invasive to an almost completely non-invasive technique. The procedure of developing of ET methodology is assessed as 4 discernible era's by Rayner (1998). The early era, which was contemporary with researchers like *Dodge*, *Dearborn* and *Huey* involved using rather crude apparatus; researches in this era, still, could lead to some important basic discoveries regarding eye movements that have stood the test of time, like average fixation time, saccade length and duration which are some of ET metrics, saccadic suppression (see section 2.6.2.1) and some theories. The second era, contemporary with persons like *Buswell*, *Tinker* and *Yarbus*, could be considered as 'a behaviourist era in psychology'. Regarding the theory development, it was not a productive era. Some researches on overt eye movement properties like eye blinks as well as some confirmatory works, are the only outcomes of this era. Perhaps Tinker's (1958) comment that 'all that has been learned about reading via eye movements has been learned' would be a good clarification for this era. The third era coincides with the cognitive revolution in psychology as well as great technological advances which led to changing eye-contingent display techniques. Also, some initial developments of theories were the consequences of this era (Just and Carpenter, 1980). And finally, the present era involves the development of sophisticated models and computer simulations for predicting the location and duration of eye gaze, required for designing interactive systems, further technological developments and a wide variety of new applications.

2.4. ET application in various domains

Original applications of ET include the three main areas of visual search, reading and scene perception (Neuroscience, 2002). Currently a wide variety of disciplines makes use of ET technique. They could briefly be identified as cognitive science, psychology, human-computer interaction (HCI), marketing research and medical research (neurological diagnosis). Specific applications include: language reading, music reading, human activity recognition, advertising and sport. Uses include: cognitive studies, human factors, computer usability, translation process research, vehicle simulators, in-vehicle research, training simulators, virtual reality, adult/infant/geriatric research, primate research, sports training, fMRI / MEG / EEG, communication systems for disabled, improved image and video communications and finally commercial applications which includes itself web usability, advertising, marketing, automotive and many others.

2.5. ET application in the geo-domain

ET was first applied in usability studies in 1947, by video recording of eye movements of pilots during landing. In the 1980s, ET was applied in HCI and then was extended into usability studies of user interfaces and websites (Chin et al., 2005).

In the 1970s, cartographers became interested in examining the map-reading process through ET in order to improve symbol design, get more control over how their maps would be perceived, and increase the overall efficiency of map use. Castner and Eastman (1984) described that studies of how, rather than where, the eye moves might prove useful in map-design experiments. They also suggested a number of ET metrics to determine the functional complexity of maps. Part II of this study (1985), uses ET to assess the holistic properties of maps. The results confirm the hypothesized correlation between subjective judgments of map complexity, termed as perceived complexity and ET data.

Chang (1985) tested the effect of experience on topographic map reading tasks by using some performance test (extraction of some absolute and relative heights) in combination with eye movement recording and interviews. The results showed that experienced readers performed better on the questionnaire test; ET data also supported that.

Brodersen (2002) in a study, applied mobile-ET combined with other techniques, for assessing the usability and design of topographic maps through giving some tasks to the participants. They examined the relationship between perceived map complexity (instructor rating, interview, verbal and non-verbal data) and ET data. The results were all promising, since the correlation between conventional methods data and ET data was high; ET and other methods, in combination, allowed deeper and more objective insight into map reading.

Çöltekin (2009), in a case study, used a combined fixed-ET and traditional usability methods to evaluate two interactive map interfaces. Overall results confirmed the hypothesis of better design of one of the two interfaces. Here also, ET data did enhance usability studies both quantitatively and qualitatively and revealed some micro level visual behavior (regarding the Identify and Redraw Map buttons' design issues) on one of the interfaces.

Alaçam (2009) conducted a combined usability testing with ET, to examine the effects of iconic representations and pop-up window usage on the usability of some web-map sites. Task performance evaluation showed there was a significant difference between the web-maps regarding their different design. Moreover, use of iconic representations, the efficiency of pop-up windows' usage and also users experience level showed to have an effect on task performance. These experiments and a few more confirm the applicability of ET in enhancing conventional methods by giving new insights to usability issues in geo-domains.

2.6. Different stages of the ET study

In conventional usability testing, user behaviours like the task duration time, task completion rate, their action and performance manner including comments, TA, emotions and their mouse hover behaviour, comprise the evaluating factors. Conducting a valid usability study, analyzing these data and combining them in a way to achieve applicable results, requires good experience and planning. Conducting a valid usability ET study is on another plane. While adding eye movement data, ET methodology typically makes use of regular usability testing, and makes it even more complicated. Unfortunately, the nature of the ET study makes it very easy to obtain fake results which surely would be worse than not doing any research at all (Pernice and Nielsen, 2009). A typical ET study usually includes different stages of data collection, clustering, analysis and interpretation. In order to achieve valid and sensible outcomes, each stage must be planned and executed properly.

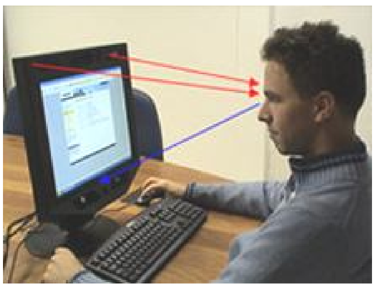
2.6.1. Phase One

The main theme of this section is data collection which refers to bringing together gaze data and other relevant data that are going to be analyzed later to make a conclusion regarding the task. General structure and function of different types of eye trackers (including head mounted and table mounted systems), their common outputs as well as the outputs of their typical complimentary usability techniques and the way these techniques are combined have been compared. Finally, the details of the data collection procedure in a typical ET experiment are provided.

2.6.1.1. Eye tracker

ET data are gathered through an eye tracker. An eye tracker determines point-of-regard (where a person is looking at) by positioning one or both eyes multiple times per second, which is used later to measure eye movements metrics (Poole and Ball, 2005).

(a)



(b)

Figure 2.2 :A representative of (a) head-mounted [ASL Mobile eye] and (b) remote [Tobii 1750] ET systems.

Two types of current eye tracker's setup include: head mounted eye tracker, which allows free head movement and is usually applied when participants will be viewing multiple surfaces or when the ability to move within a restricted area is required (drivers, pilots, athlete practicing and also paper prototype or other out of the box studies) and stationed (other terms are fixed, table mounted or remote) ET systems which can work in a limited area like the computer screen or document based studies (figure 2.3-b), and some other applications which do not require the user to move e.g. shelves studies and marketing (figure 2.3-a). Since their related software is more developed, it would be more efficient if fixed systems are applied for screen-based applications. There are two types of remote systems which may allow a person to move his head partially in a restricted area or not at all. A completely stationary eye tracker which also typically requires a chin rest or bite bar, has a higher accuracy. An eye tracker measures the rotation of the eye with respect to the measuring system. If the measuring system is head mounted then eye-in-head angles are measured, and if it is a fixed system then gaze angles are measured. Most eye trackers use a sampling rate of at least 30 Hz until 50/60 Hz. Likewise several video-based eye trackers (see section 2.6.1.3) run at 240, 350 or 1000/1250 Hz to capture very rapid eye movement details during reading or neurology studies.

(a) (b)

Figure 2.3: Illustration of using stationed system for (a) marketing application [Tobii 120] (b) document based application/reading [SMI-RED250]

Most ET systems today work by reflecting IR light toward the eye and recording the reflection patterns with some cameras. Then the point of gaze is calculated for the reflection data by a geometrical model. Different eye trackers differ in their design and data collection algorithms (which are methods of analysis to determine the location of fixations); however, most head-mounted systems include two cameras: one for recording the field of vision from the subject's perspective and the other for tracking eye movements via IR reflections. Participant's image of point-of-regard (as a dot) then would be superimposed on top of the image of the field of vision (figure 2.9-b). These two images are saved as alternate frames by a video recorder, which means the output format of a mobile eye tracker would be a video recording of combined images.

Stationary systems on the other hand, use some cameras which either are external or placed in the frame of the eye tracker screen and points directly at the participant's eye (Poole and Ball, 2005). Similar to mobile systems, here also the eye gaze data as a moving point is superimposed on the screen logged stimulus (figure 2.6), which could also be exported and saved as video clips (for instance in avi format) if required.

The reason for a possibility of free head motion in head mounted systems is that these systems measure the pupil glint from multiple angles in order to differentiate eye movement from head movement. In the same way, some remote systems recently use several smaller fixed sensors placed in the computer monitor frame such that the glint underneath the pupil could be measured from multiple angles which provides the chance of head movement within around one cubic foot (Cooke, 2005).

Apart from data collection software, the main components of a mobile eye tracker are: glasses mounted optics with built-in ET cameras which may track one or both eyes, IR light emitting diode (LED) which is placed close to the eye camera for emitting IR light to the eye, a head mounted scene camera, a recorder (VCR) either as a built-in piece of the head mounted part or as a separate portable component and a processor computer with a high storage disk space (figure 3.2). The head mounted part and the external recorder are required to connect to each other during the test. Depending on the system type, the recorder and the PC may or may not connect to each other during the test for remote control applications. The remote system on the other hand is composed of the eye tracker itself (in the form of a flat monitor) with special built-in autofocus cameras and LED in the frame of the screen or external cameras close to the monitor, microphone (to combine ET with TA), keyboard and mouse (for using the under processed application) and a processor computer (figure 2.2-b, figure 2.4, figure 3.9). If it is desired to minimize the amount of equipment needed to implement the project in someplace which aids in research like a home or school, the single computer configuration is preferable; however, in an ET lab usually a two-computer setup is applied (figure 2.4). The procedure is that the stimulus is run on an application computer, also the stimulus with its superimposed gaze data is run on another host computer for the real time monitoring. The application computer can generate signals in response to user generated events like a button press or mouse movement; the host computer can also generate and send signals to start or stop tracking.

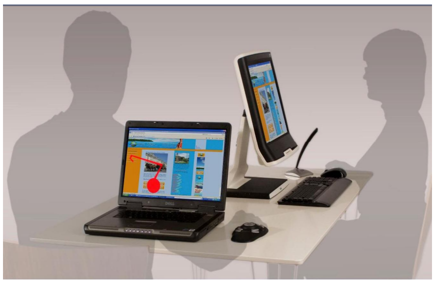


Figure 2.4: Using a two-computer setup for real time monitoring during an ET experiment (URL2).

As mentioned before, the mobile and stationed systems have different structures and configurations; likewise they have different potentials and limitations. Selecting a mobile system or a stationed one is determined by task requirements and the budget. Each of them is most suited for a particular application. There are also variations in system specifications and data quality within each type. Although a few attempts regarding comparing different systems within the same experiment has been made, e.g. (Nevalainen and Sajaniemi, 2004), all of them are restricted into evaluating just two or three systems; this makes it impossible to generalize their finding and come to a conclusion. The followings are some comparisons regarding the general potentials and limitations between mobile and stationed systems:

- Calibrating mobile ET systems for people with glasses or corneal irregularity is not so easy, and for test conditions that visual impairment is correlated with other variables, it may leads to a biased sample (Mayr et al., 2009). Furthermore, some systems perform a poor calibration in uncontrolled environmental conditions (Hansen and Hammoud, 2006). In the same way, remote systems have some limitations for calibration; during ET with a stationed system, head movements may cause a delay until the eye tracker can reacquire the eye; this may also cause a loss of calibration (Namahn, 2001).

- Compared to mobile systems, fixed systems are less obtrusive. Utilizing a mobile eye tracker implies the subjects wearing goggles which obstruct a small portion of their visual field and knowing that their gazes are being tracked; this unusual feeling may bother the subjects (Mayr et al., 2009).

- In contrast to fixed systems, a mobile eye tracker has a limited temporal accuracy, which means some short fixations can be easily missed. Also, ET works best if the eye tracker is calibrated to a specific fixation distance. However, due to free head and body movement in mobile systems, this distance is not constant; consequently, their spatial accuracy is lower than fixed systems (Mayr et al., 2009).

- Although head restriction is a drawback of remote systems, due to their relatively high durability and low cost compared to head mounted systems, they are more popular than head mounted ones (Jacob and Karn, 2003).

- Static ET systems usually are provided with the software for automatic data analysis. In mobile systems however, due to the inter-individually variable behavior of eye movements as well as constantly changeable background which leads to changing size and position of different ROIs, automatic analysis is limited. As a solution, many researches use only inter-individually similar short tasks with an expected eye movements' range, which of course limits the generalization of ET results. Also, some recent developments (e.g. ASL's GazeMap software) allow for detecting dynamic lookzones by applying edge detection and object recognition algorithms which also facilitate multiple subject analysis (Pernice and Nielsen, 2009).

2.6.1.2. Eye tracker's output

What an eye tracker collects as raw data is a stream of 2D points (with usual sampling rate of 50 to 250 Hz) or points-of-regard, which are points in space where a person is looking. They are usually used in ET studies to reveal where visual attention is directed at (Poole and Ball, 2005); some systems provide the torsional movements or Z in the data file as well. Depending on tracker's type these information could be provided for either one or both eyes. Data collection is handled by ET software. All ET software share common features. Software catalogs ET data is in one of two ways: data are stored in video format (using a small dot to represent eye movement) or data are stored as a series of X,Y coordinates related to specific grid points on the computer screen (Cooke, 2005); in video recorded data format, data collection software extracts 2D gaze data points through image processing of video frames. To give these points meaning, ET analyzing software later should map the points onto the meaningful objects at which the eye was looking (Reeder et al., 2001).

Apart from gaze location, time and duration of samples, there are also other aspects of ocular-motor performance, provided by some ET systems, such as pupil dilation and blinking rate as well as calibration-related data like eye position, distance from eye tracker to eye and validity code. Other possible non gaze-related outputs like stimulus recording (e.g. a web page transition and scrolling information), scene camera, timestamps, key strokes, mouse clicks, and user camera and audio which are automatically synchronized to gaze-related data in order to provide a complete view of the participant's behaviour during the test.

2.6.1.3. Data collection operation

To date, many different ET technologies have been used to measure point-of-regard of the eye. An old technique for instance was electro-oculographic which used some electrodes mounted on the skin around the eye to detect eye movements by measuring electric potential differences. Another invasive technique used some kinds of contact lenses with a metal coil embedded around their edge; it could locate eye movements through measuring the electromagnetic field created by moving metal coil along with the eyes. Infrared oculography or IROG (versus video oculography or VOG) was another technique which utilized the IR reflection for detecting eye movements. IR reflection of some LED is received by two nasally and temporally located phototransistors and then transformed into voltages; the result of subtracting these two voltages would be proportional to the angular deviation of the eye (Špakov, 2008). Today, however, eye recorded video images, created through measuring the reflection of IR light that is shone onto the eye are mostly used.

Modern eye trackers may use different features of the eye like the iris-sclera boundary, corneal reflections, and the apparent pupil shape to extract point-of-regard (Poole and Ball, 2005). Video-based eye trackers usually measure point-of-regard by the corneal-reflection/pupil-centre method. These eye trackers need to be fine-tuned to each individual's eye movements by a *calibration* procedure that 'teaches the ETsoftware how the eyes of a particular subject look and behave at certain screen positions' (Manhartsberger and Zellhofer, 2005) and consists of collecting point-of-regard data for a set of predefined target points in the visual display or in the real world. Some eye trackers are capable of saving calibration data for each person and reuse it later without recalibration. Usually a dot displays on the screen and then the eye should fix for a short time on it, then the system records separately this pupil-centre/corneal-reflection relationship for each specific x,y coordinate on the screen. This procedure requires some specific hardware and software. ET hardware is either mounted on a user's head or mounted remotely. Both systems measure the corneal reflection of a LED, which illuminates and generates a reflection off the surface of the eye. This makes the pupil appear like a bright disk and also creates a small glint below it (Figure 2.5). When the software could recognize the location of the glint and the centre of the pupil, the vector between them is measured. Then with some trigonometric calculations, point-of-regard would be obtained (Poole and Ball, 2005).

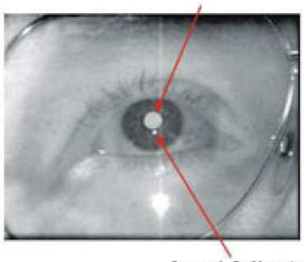


Figure 2.5: Eye tracker traces eye movements by recording the location of the glint and the pupil, using IR light.

Since IR is not part of the visible light range, its reflection does not disturb the eye. Another advantage of using IR is that the surrounding light does not affect infrared detectors; only during calibration, due to dimness of the glint, environmental light sources need to be controlled (Graf and Krueger, 1989).

Another related concept is the use of *dark pupil* or *off-axis illumination* versus *bright pupil* or *coaxial illumination* tracking method in different ET systems; the bright pupil method aims the IR light beam from the camera directly towards the eye such that it reflects back towards camera, while the dark pupil tracking method aims the IR light at the eye from outside of the camera's optical axis (DeSantis et al., 2005). Dark pupil and bright pupil tracking methods works better for people with dark and blue eyes respectively; each method has its own benefits, for instance, the optics for the bright pupil method work better in dim light, create fewer track losses and less noise, but it is also more sensitive to abrupt changes in ambient light (Merchant, 2001). Currently these techniques are combined to eliminate the need for two separate eye trackers.

2.6.1.4. A typical test procedure

Before deciding to conduct an ET experiment it is wise to clarify the scope of desired evaluation and to justify the use of ET by the process-related questions of the study or a list of question-hypotheses that ET is expected to answer (Goldberg, 2003). The way an ET task performance is planned has a high influence on the resulting data later. Before the experiment, the devices must be checked carefully to make sure they work properly; the test environment should be arranged to be similar to the real conditions of using the application. In a typical test protocol, just like any other usability testing, the ET experiment presentation should be piloted. After choosing the correct number and category of participants (they are usually end-users; recruiting users with proper eye for calibration, although bias the results, raise the accuracy) according to the purpose of study, there is usually a pre-questionnaire for the user's demographic data, a test of the subject's eyes for calibration, a task performance session, and a post- questionnaire to obtain the user's feedback of the test. Depending on the final deliverable of the study and in order to obtain deeper knowledge of user's motivations and thoughts while carrying a task, ET may be combined with concurrent TA or retrospective thinking aloud (RTA). RTA is also recognized as post experience ET protocol or PEEP; the difference is that here user is asked to apply the dynamic replay of their eye movement trace as a cue to encourage RTA after the test. It is proved that verbalizing in TA during ET usability testing is affecting eye movements. For some product like a *heatmap* and *gaze plot,* which are two ET software products (see section 2.6.3.2), it is recommended not to ask users to TA. Visual distractions like moving or colourful objects can modify the expected position of gaze; it is better to reduce such factors in the testing environment or on the screen. Usually in usability testing, asking user's opinion about a prototype rather than having him to do tasks is considered a poor usability methodology. The same fact should be considered in applying ET technique. The words and gestures the experimenter use to prepare the participant could have great impact on how he will proceed throughout the study, such as training participants to talk about their actions without solving problems, suggesting solutions or studying the prototype. It is important to make the test persons aware of the fact that it is the prototype or the method that is going to be tested, not their ability (Pernice and Nielsen, 2009). Another important rule is that test persons should be unfamiliar with the product and be unaware of the specific purpose of the research (Cuddihy et al., 2005). In order to define some specific tasks script for the test, those tasks which may have problems from what typical end-users typically do are identified. These tasks then are simply written down, such that they do not instruct users completely how to do them. To avoid learning effects, the order of the given tasks should vary.

Apart from general required considerations in a typical ET test, there are some requirements which are specific for each tracker's type. Since mobile systems are mostly applied outdoors, employing them entails some environmental consideration including weather (as a distracting factors e.g. rain), light condition and humidity (as a calibration limiting factor which also applies for fixed systems), disturbance by curious people (better to avoid rush hour), and also the battery charging duration. Likewise, using fixed systems necessitate some considerations; they include improving calibration by using a stationary chair for user and asking him not to move much as well as reducing blinking rate of a dry eye which stared at the monitor by reducing test duration less than 90 minutes. Also to improve the test accuracy the experimenter needs to take control of the mouse to stop the task, he also should sit slightly behind user in order to not to encourage conversation, reduce user movements by delivering questions towards him and taking notes of problematic areas to focus later (Pernice and Nielsen, 2009).

2.6.1.5. The number of required participants

According to Goldberg (2003), the number of required participants in a usability testing is determined by some factors including: expected effect sizes, statistical power and expectations within a domain. Considering statistical power, ET creates a large number of data. However, in the case of significant aggregation and averaging of ET data prior to statistical analysis, the number of test persons may need to be increased. For instance, a user may create 1000 fixations over the stimulus under evaluation. During the pre-statistical analysis, the 1000 fixations turn into average gaze duration within 10 of ROI (see section 2.6.2.2- AOI algorithm). Pilot studies also would increase the number of participants

Having a lot of ET and other usability experiments, Nielsen has come up with some numbers for ET participantes (Pernice and Nielsen, 2009). To conduct a qualitative usability testing, he suggests around 5 persons. However, for an ET usability testing, before deciding on the number of users, the main deliverable of the study must be defined. If a live test behaviour along with gaze replays is running as primary data, then 5+1 person (to account for lost or poor data) is needed; in case of making good notes during the live testing, he suggests even 3 acceptable ET recordings would be enough. Also, if the main deliverable is based on the quantitative statistically significant data of a heatmap which is one of the ET outputs (see section 2.6.3.2), then the data from 30 users per heatmap is required (39 users to account for data loss due to poor ET). The variability in heatmaps, as a result of micro-level ET data quantification, depends on the number of users; therefore, he recommends using heatmaps just as illustrations to confirm other findings and not as a statistically significant primary data (Pernice and Nielsen, 2009).

2.6.1.6. Combining ET with other usability techniques

Depending on what a user has in mind, there might be several dissimilar reasons for a unique eye movement pattern, e.g. a fixation on a face in a picture can be a sign of like, dislike or recognition. Likewise, a fixation on an interface display may indicate interest or inefficient search. Due to this issue, ET has a limited applicability in isolation and it requires the support of some other explanatory techniques. ET is often used along with other conventional usability evaluation techniques in user experience research; these two techniques are complementary to each other and the combined data from them provide a broad overview of the problems a user may run into while performing a task in some interface, and increase the validity of interpreting user behaviour (Bojko and Schumacher, 2008).

Questionnaire, TA, RTA, interview and experimenter’s observations of the user are the usual accompanying techniques of ET which are applied before, during and after an ET session. There is not a common agreement on using TA or RTA in literature; while RTA could prevent interrupting the user (Ball et al., 2006) during a task, which is a huge benefit, there is a danger of memory loss due to the created time delay. Apart from the argument of ‘what qualitative techniques are preferable to be combined with ET’ and ‘which combination order provides best outcomes’, there was always the technical issue of effective synchronization of these techniques. This technical obstacle has been attended to in new system developments by a *dynamic display*. Recent fixed system analyzing software provide the opportunity of integrating ET data and other data from the operating system like key strokes, mouse clicks, document scrolling, window sizing and web page URL recording, along with other user data like their voice, facial expression and behaviour through audio and video recording during and even after (for simultaneous recording of RTA over stimulus replay) the test. All these data along with the test set-up information reside in a built-in database that could be used later to extract relevant data for some particular processing (figure 2.6). In mobile systems, however, these facilities are limited to integration of superimposed gaze data over stimulus with the user voice and pupil size.

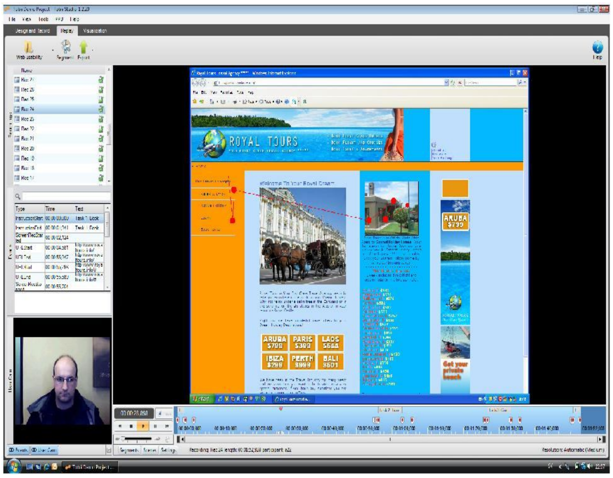


Figure 2.6: Illustration of an ET analysis software ’Tobii Stadio ™’. It provides synchronous user data including: participant’s audio and video recording, and screen logging with superimposed gaze point (URL2).

Real time display which provides the opportunity of observing the test session including subject’s screen with the ET data superimposed (as raw data or fixations) and all other input data of the subject, is another advancement of modern ET software which is perhaps relevant to be mentioned here. This functionality which gives real time insight into the subject’s behavior is utilized for instance in moderating the test, presenting the test to clients, real time reaction on subject’s behavior and the assurance of recording quality (Manhartsberger and Zellhofer, 2005).

2.6.2. Phase Two

The second phase of an ET study is devoted to data clustering which illustrates the process of altering eye gaze data to classified categories of fixation and saccades. Furthermore, this section provides some general information regarding different eye movements, common ET metrics and the ones often used in different geo-applications and also a brief introduction to the common clustering algorithms.

2.6.2.1. Eye movements and ET metrics

Natural movements of the eye can be classified as five different types including *smooth persuit, optokinetic nystgmus, vestibulo ocular reflex, vergence* and *saccades* (Purves et al., 2004). The most common eye movement we make is the *saccadic eye movements*, which includes saccades and fixations. Fixation (usually lasts between 250 and 500 ms) occurs when the eyes are relatively stationary and focused on some specific location of a stimulus; saccades (typically last between 25 and 100 ms) on the other hand, are rapid jumps of the eye from one place of the stimulus to another. Information processing known as *decoding* happens only during fixations (Bach-y-Rita et al., 1971). Decoding lasts for a while and then another saccade will occur. The eyes are almost blind during saccades; this fact is identified as *saccadic suppression*. Due to the short duration of saccades however, the blurring goes mostly unnoticed (Bojko and Schumacher, 2008). Furthermore, it should be considered that saccadic eye movements could not just be divided into points (fixation) and lines (saccade). While during a processing, the eye is concentrated on some object, there are still some smaller eye movements which can be later mapped into one point, meaning that as long as the point-of-regard is below some threshold of *area, duration* and *velocity*, it would be a fixation. In the same way, during movement from one fixated point to another, the direction of point-of-regard is not always a straight line (figure 2.7). For the analysis purposes however, the actual path travelled during saccades is mapped into a straight line, since it does not affect most typical ET studies (Salvucci and Goldberg, 2000).

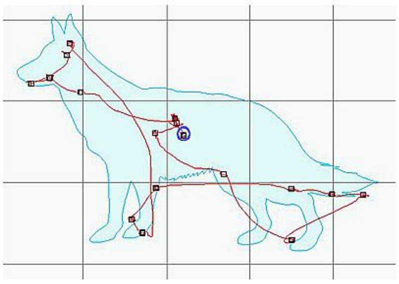


Figure 2.7: Illustration of a pure scanpath with fixation (URL4)

While the eye is fixating on some feature, the visual field is split into two main parts as foveal and peripheral. The foveal vision (including the central two degrees of the visual angle) provides the ability of receiving detailed visual information; while the input from wider eccentricities through peripheral vision is not so informative (Manhartsberger and Zellhofer, 2005). ‘Since the saccade is ballistic, its destination must be selected before movement begins; since the destination typically lies outside the fovea, it must be selected by lower acuity peripheral vision’ (Jacob, 1995), which means the direction of the gaze is guided by the peripheral perception. Furthermore, what an eye tracker collects as point-of-regard is just from foveal region and this limits the interpretation of eye movement based on gaze landing points.

Fixations and saccades are the main eye movement *measurements*, and a number of derived *metrics* like *fixations duration* or *number of saccades* are based upon them. Some other metrics like *spatial density of gaze point* may drive directly from point-of-regard raw data. Further analysis of fixations can detect which predefined or non-predefined objects were captured by each fixation and then higher level metrics like sequences and transitions among these objects can be calculated (Goldberg, 2003). ET metrics could be either temporal, which are related to the time duration spent on the stimulus or be spatial, related to their degree of coverage of the stimulus. Other measurements of eye movement are *scanpath* and *gaze*. A series of fixations and saccades forms a scanpath; fixation locations in a scanpath are the representative of the location of processed information. An optimal scanpath in a search task is the straight line towards the target, with relatively short fixation duration at the target. The sum of all fixation durations within a defined area is known as gaze (other terms are *dwell, glance, fixation cycle*), a fixation occurring outside a ROI means the end of that gaze; gaze may also include the spent time for the short saccades between these fixations (Goldberg and Kotval, 1999).

The duration of a saccade or fixation depends on the defined task, related application as well as human cognitive state. For a linguistic text reading for instance, fixations duration (200 ms) is different from viewing of a scene (350 ms); a saccade duration on average could be estimated as 200 ms as well. Non-gaze related metrics like pupil diameter, blinking number and frequency and the latency before a blink, as an index of cognitive workload are also measured by some systems. The increased and decreased blink rates are proved to be indicative of fatigue and higher workload respectively (Bruneau et al., 2002). Wider pupil size can also point to more cognitive effort (Marshall, 2000), however these metrics can be affected by many other factors and are not accurate enough to be relied on. For instance, any change in luminance could cause a pupil diameter's change.

Different attempts during the procedure of evolving of ET methodology have led to a heterogeneity regarding the terminology and definition of ET measurements (e.g. the threshold for a fixation) and the underling concepts for usability analysis (e.g. relating a long fixation to difficulty or interest). In order to be able to compare different researches and extract meaningful information from eye data, ET metrics are required to be standardized. Another issue regarding defining the ET metrics is to select metrics with proper types and quality, corresponding to the type of the study and the questions ET is going to answer. In order to identify relevant metrics for a particular task, Jacob and Karn (2003) recommend to find out about those aspects of eye position which may explain the usability issue. They also presented the six most common ET metrics applied in 24 different HCI researches including total number of fixations, gaze on each ROI, total fixation duration mean, number of fixation on each area of interest, gaze duration mean on each are of interest and total fixation rate. karn (1999) also presented a list of simple to more complex metrics which typically are assessed in data analysis in sequential order; they include: 1. fixation, saccade, and pursuit of eye movement 2. scanpath, overall fixation duration within ROI and the matrix of transition probabilities between ROI 3. scanpath shape, complexity and variability. For similar purposes and by summarizing earlier work, (Ehmke and Wilson, 2007) came up with 28 ET metrics and linked the metrics to their related cognitive processes or usability problems. A summary of these metrics along with an indication of whether they are already applied in the geo-domain based on experiences found in the literature review is provided in table 2.3.

Table 2.3: Applied ET metrics and related cognitive processes or usability problems in general (Ehmke and Wilson, 2007) and in geo-domains.

ET metrics	Cognitive process or usability problem	Applied in geo-domain
Fixation-related		
Time to first fixation on target	Good (if short) or bad (if long) attention getting properties	yes
Fixation spatial density	Focussed efficient searching OR widespread inefficient search	
Fixation duration (Fixation length)	Difficulty in extracting information OR more engaging; voluntary (>320 ms) and involuntary (<240 ms) fixations; needs further investigation	yes
Fixations on target divided by total number of fixations	Low search efficiency	
Number of overall fixations (Fixation count)	Less efficient search due to sub optimal layout	yes
Repeat fixations (post-target fixation)	Lack of meaningfulness or visibility	
Fixations per ROI	Element/area more noticeable OR element/area more important	
Percentage of participants fixating on ROI	Attention-getting properties of an interface element	
Fixations per ROI adjusted for text length	Element harder to recognise	
Saccade/fixation ratio	More processing or less searching	
Saccade-related		
Number of saccades	More searching if more saccades	
Saccades revealing marked directional shifts	User's goals changed OR interface layout does not match user's expectations	
Saccade amplitude	Meaningful visual clues if larger saccades	
Regressive saccades (backtracks/regressions)	No meaningful visual clues, changes in goals, mismatch between users' expectation and the observed interface layout	
Saccade duration	Low image quality such as blurred or low contrast	
Scanpath-related		
Scanpath duration	Less efficient scanning (if long)	
Scanpath direction	Indication of search strategy	
Scanpath length	Less efficient searching (if long)	
Small spatial density of scanpath	More direct search	
Scanpath regularity	Search problems due to lack of training or interface layout problems	
Transition matrix (back and forth between areas)	Uncertainty in search OR search order efficient and direct	
Transition probability between ROIs	Efficiency of arrangements of elements in user interface	
Gaze-related		
Gaze (dwell)	Measure of anticipation OR attention distribution between targets	yes
Gaze orientation	Feedback about success of design features	
Gaze duration on ROI	Difficulty extracting or interpreting information from element	yes
Number of gaze per ROI	Possible importance of element	yes
Spatial coverage calculated with convex hull area	Scanning in a localised or larger area	

2.6.2.2. Clustering eye movement data

Raw eye movement data are huge and messy. Analyzing these massive data if not impossible is a tedious and highly complex process. Adding data from different subjects together for the purpose of obtaining an overall result makes this procedure even more complex. *Clustering* is a solution for the analysis of massive ET data by reducing the data volume such that the main characteristics of data, required for understanding cognitive and visual behavior, are retained. Usually it separates and labels raw ET data as fixations and saccades (i.e. fixation identification algorithms), for further processing, later these clustered data may cluster again and create more condense results (*fixation clustering algorithms*).

Clustering process is an essential part of an ET study that can later, have dramatic effects on analyses and interpreting the data (Karsh and Breitenbach, 1983). The outputs of different clustering methods are not identical; furthermore, by changing parameters' thresholds the same method may produce different outputs (Shic and Chawarska, 2008). The quality of input data (device accuracy) is also a distributing factor in the clustering's output. Theoretically there are many available algorithms for clustering (Duda et al., 2001), however in practice, only a few are applicable to classify highly volume gaze data. The following provides some knowledge regarding typical applicable clustering methods for processing gaze data.

Clustering raw data points is executed through some forms of fixation identification algorithms. Such algorithms ‘remove raw saccade data points and collapsing raw fixation points into a single representative tuple’ (Salvucci and Goldberg, 2000). Data reduction has two parts; one part relates to merging close eye movements by collapsing them into the related fixation point as well as omission of smaller unwanted movements which cause noise during fixation (Ditchburn, 1980), and the other part is about reducing saccadic data points by exclusion of the actual path between two fixations (figure 2.8). Salvucci and Goldberg (2000) classify common fixation identification algorithms by five representatives (VT, HMM, DT, MST,AOI), based on measures like accuracy, speed, robustness, ease of implementation and parameter space and with respect to spatial and temporal characteristics. Most of clustering algorithms, applied in available ET tools, are variation of their taxonomy with further filtering methods for noise reduction and validity inspection. Moreover, although these methods are usually applied for offline data processing, the same logic with different implementation can be used in real time fixation detection algorithms (Špakov, 2008).

Based on their taxonomy, the three algorithms which make use of temporal information (HMM, DT, MST) provide more accurate and robust fixation identification. By using local adaptivity, interpretation of temporally adjacent points will influence the interpretation of a given data point and this, will result in reducing noise. The pseudo codes for these algorithms are provided in appendix A.

1. <u>Velocity threshold:</u> This algorithm first calculates point-to-point velocities for each point based on the distance between that point and the next point in the protocol. Then it classifies each point as a fixation or saccade based on a velocity threshold. As the next step, it collapses consecutive fixation points in a group and discards the remaining saccades. Finally, it maps each group to a representative fixation at the centroid of its points. The problem with this algorithm is that it requires a threshold value and it is not robust with the noise.

2. <u>Hidden Markov model:</u> Similar to velocity threshold, this algorithm also calculates velocities between each subsequent point pairs. Then it decodes velocities, with a two state HMM in order to differentiate fixations and saccades. HMMs are probabilistic finite state machines which utilize probabilistic analysis to find the most likely identifications of each point in a protocol through a process of decoding and using dynamic programming. The applied algorithm here is a two-state HMM in which the states represent the velocity distributions for saccade and fixation points. Decoding is the process of optimal assigning of each points to one of the two states which maximizes the probability of the protocol given the HMM. The rest is similar to VT, which means the consecutive fixations are collapsed in a group and are represented by a centroid point. Since it utilizes the probabilistic analysis to calculate velocities, this method is more robust than VT.

3. <u>Dispersion threshold:</u> This algorithm considers the spatially close consecutive points that are within a dispersion threshold as a fixation. It also incorporates a fixation duration threshold to lessen the equipment variability. Based on Widdel’s (1984) data reduction algorithm the DT structure is as follows: a moving window calculates the dispersion value for its interior points and compares it with the threshold, while it spans data points. The window is initialized over first points of the protocol and selects a minimum number of points, based on a predefined duration threshold and sampling frequency. If the calculated value was more than threshold, no fixation is represented and the window removes first point from interior points and advances forward by one point. If it is less than threshold, then the window represent a fixation. Here, the window is expanded to include next points until the window’s dispersion value is more than threshold. After that a fixation at the centroid of this final window is recorded and the window

would be removed from the points. This process of moving window to the right continues till it gets to the end of the protocol. Dispersion-based algorithms are sometimes used to locate clusters within minimum spanning tree network representations.

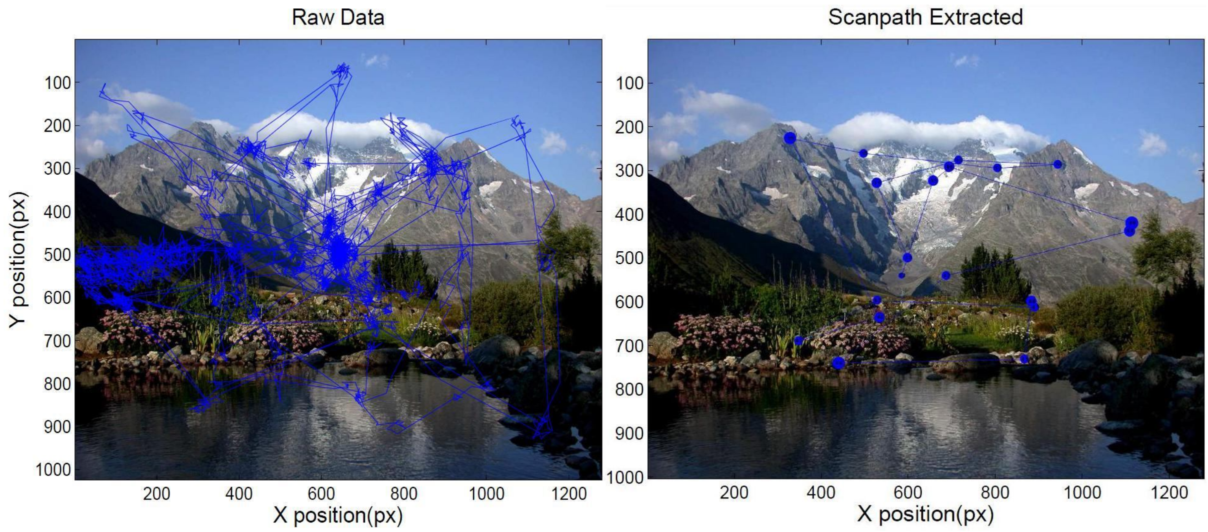
4. Minimum spanning trees: This algorithm is based on the minimum spanning trees; it is a tree that makes connection between the collected gaze data points such that the total length of line segments of the tree is minimized. It usually used for off-line analysis. 'the main idea of the algorithm is to use one of the methods of data search optimization to find pair of adjacent samples, which are separated by a distance greater than a given threshold, thus identifying saccades' (Špakov, 2008).

5. Area of interest: This algorithm is applied to identify only fixations that happen within some predefined target areas. It requires the definition of these AOIs and usually used for off-line analysis to keep identified fixations close to the relevant AOIs. The algorithm starts by labelling each point as a fixation point for the AOI in which it lies, and otherwise labelling them as saccade points. Then it collapses the consecutive fixation points belong the same AOI into fixation groups and removes saccades. Next, it removes fixation groups that are under a predefined duration threshold. Finally, it transforms each remaining fixation group to a representative fixation at the centroid of its points.

Also, effective data-driven clustering algorithms can be studies in two groups of distance threshold and mean shift:

1. Distance threshold: This algorithm classifies fixations in different clusters based on their distance from each other. Once the distance between two fixations is less than a predefined value, they belong to the same cluster. The algorithm begins by looking at a fixation which may already assigned or not assigned to cluster (and assign it to a new cluster if it is not processed yet) and then calculates its distance to other fixations (which did not analysed yet). If the calculated distance was less than the threshold, they would be added to the same cluster and also if the added point where already part of another cluster, then these clusters would be merged (Špakov, 2008).

2. Mean shift: This algorithm may have two implementations. The main difference is that one of them considers the sequence occurring fixation in a cluster and the other does not. Mean shift makes use of the distance threshold, but it also includes a pre-processing stage. The algorithm begins with shifting fixations into denser locations, till they can be easily separated into clusters; the shifted value would be the weighted mean of the surrounding points which is based on a Gaussian function. Then, it applies the distance threshold algorithm to these points to extract clusters and finally outliers (clusters with small total fixation duration) are discarded. (Špakov, 2008).



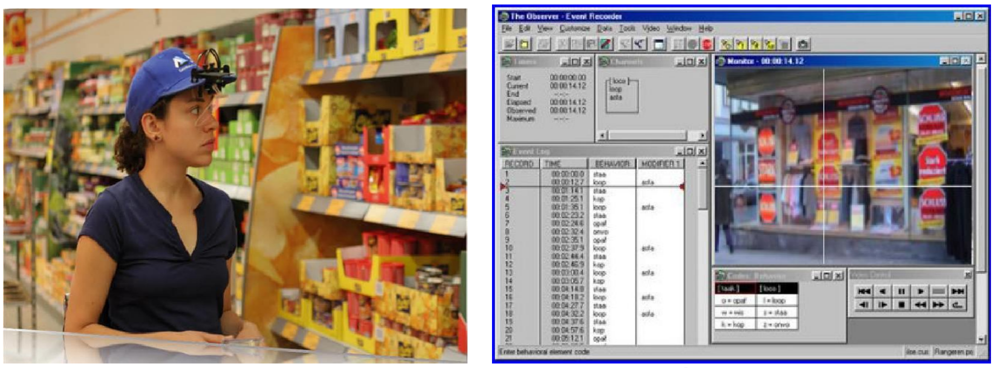
(a) (b)

Figure 2.8: A representative of raw scan data compared to a scanpath: (a) The scan of point-of-regard (raw) data, (b) The same clustered data with fixations and saccades (Torstling, 2007).

Recent ET software have the facility to extract fixations from gaze data for their later analysis as well as for the export functions. Their built-in clustering algorithm is rather adjustable; however, in order to achieve deeper analysis and more flexibility for specific applications, it is possible to export raw gaze data or clustered data to other software.

Filtering may also refer to eliminating irrelevant gaze data created as a result of eye tracker malfunctioning or uncontrolled and non-task relevant eye movements; these types of filtering require task related logical judgment and human cooperation and cannot be done just automatically. In order to raise accuracy and validity of raw ET data, other filtering algorithms have been implemented in new software; *ClearView* provided by Tobii eye tracker for instance, is capable of filtering data based on the *validity code*, to only present data with a high degree of certainty. *Eye filter* is its another filter which is applied to increase longevity of calibrations and get higher accuracy by utilizing the data from the average of the two eyes; eye filter is capable of filtering eye data based on recording conditions which could be either binocular or monocular (URL2).

In mobile ET due to the possible free head and body movements, the background changes constantly (figure 2.9). This means even during a fixation of the eye on some object, the location of point-of-regard is not constant. To calculate fixation in this case, the software is required to recognize the objects on the video frame and combine this information with ET data. This process which requires development of sophisticated feature extraction object processing algorithms for the mapping of fixation points to visual stimuli, is not fully possible yet in present ET software (Mayr et al., 2009) (figure 2.10).



(a) (b)

Figure 2.9: 'iView X™ HED' mobile ET application in marketing, (a) changing of background due to the free head and body movements of the customer, (b) related video analysis package 'NOLDUS Observer Video-Pro™' for statistical analysis (URL8).

Figure 2.10: ASL Software solution (gaze tracker) for extracting ROIs and mapping of fixation points to dynamic visual stimuli (URL1).

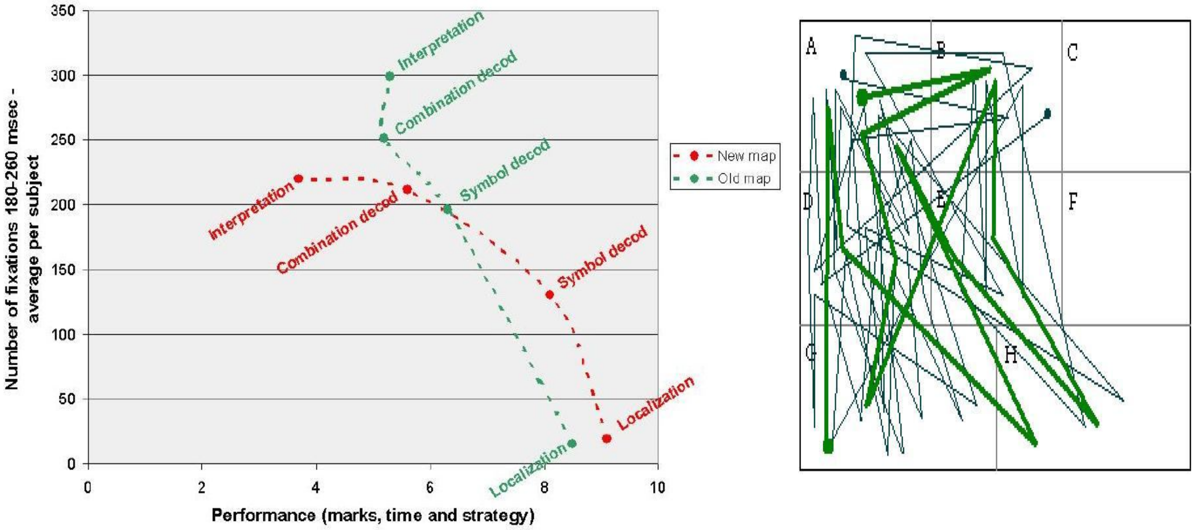
2.6.3. Phase Three

Data analysis is the third stage of an ET study which may be carried out either with the assistance of ET software or by developing personal tools. Usually before analysis, the resulted scanpath of data clustering is briefly reviewed through some visualization tool to check the logical matching of patterns with the stimulus. Then a ROI detection algorithm, with further automatic or manual error correction from the observed image would be run (URL3) (Špakov, 2008).

2.6.3.1. Analyzing eye movement data

Data clustering in some literature is regarded as part of data analysis. However, it is separated here as a pre-processing stage which prepares raw data for the analysis purpose by transforming them into fixations and saccades; these fixations could be clustered again to give a deeper insight about overall scanpath distribution around the stimulus. Patterning of clustered data or discovering similarities among scanpaths (figure 2.11-b) is part of data analysis which can reveal cognitive strategies that drive eye movements (West et al., 2006). Patterning of scanpaths is done via some algorithms.

Depending on task requirements, data analysis may also include either comparing sensible ET metrics against each other (e.g. number of fixations per fixation duration) or comparing ET metrics between subjects (e.g. number of fixations or mean fixation duration) or for different tasks within subjects (e.g. number of fixations or mean fixation duration). Data analysis could also refer to associating related ET metrics to the outcomes of traditional techniques like task completion time to see whether they support each other or not. For instance the analysis of ET data versus conventional techniques, represented in figure 2.11-a, would be that there is a linear relation between performance and the number of fixations, in two different maps, while the tasks are getting more difficult. Also, by increasing the task complexity the number of fixations increases.



(a) (b)

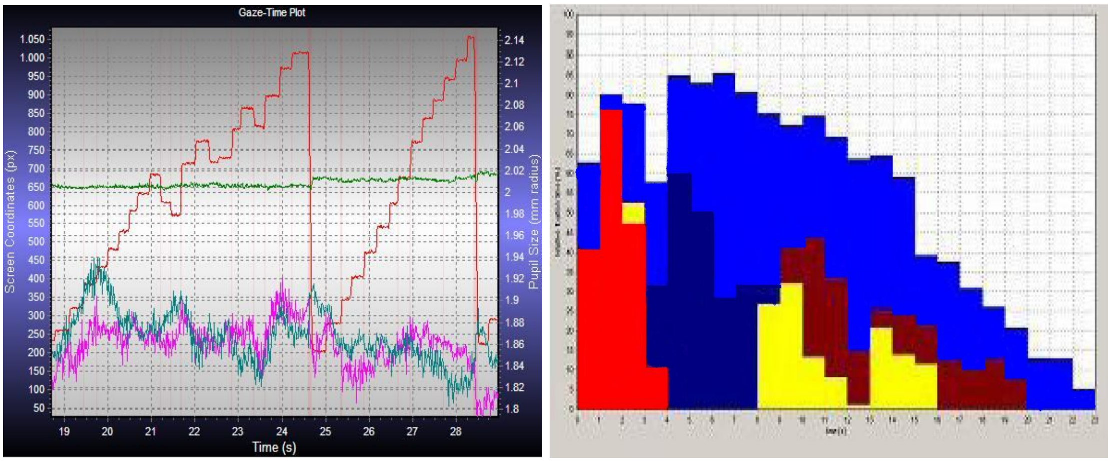
Figure 2.11: Examples of gaze data analysis: (a) ‘fixation versus performance’ of two different maps, as the tasks are getting more difficult (Brodersen et al., 2002), (b) Patterning of scanpaths; original scanpaths represented as thin green line and compressed scanpath as a thick green line (URL12).

Data analysis can be carried out through developing personal tools as well as using ET software. Recent ET software cover calculating the essential ET metrics; their embedded statistical tools allow for displaying gaze statistics or graphs, based on calculating ET metrics such as time to first fixation, fixation count, fixation duration, gaze time distribution, average gaze time and transition matrix. Also some mouse click related statistics like time from fixation to click can be calculated by the software. Using ET software makes data analysis easier and much more efficient, still depending on the required analysis and related metrics, in some cases it is necessary to export raw data to other software or develop personal tools.

2.6.3.2. Visualizing statistical outputs

Visualization techniques cause information perception and grasping statistical data to occur most efficiently. In addition to visualizing ET statistical data in the form of charts for the interpretation purposes, there is also the possibility of using visual statistical products such as heatmaps and gaze plots. Such products are created by most ET softwares and like other statistical usability outcomes, when created, need to be combined to conventional techniques to be clarified and interpreted. Some products like cluster, 3D visualization and bee swarm are less common. The followings are the description of some of ET visual statistical outcomes:

- *One-dimensional visualization*, which is employed in some ET software, represents X or Y plotting of gaze data points (or other data like pupil dilation, eye movement velocity and acceleration or even higher level data like scanpath) of one or more users against time. Sometimes, different types of data of the same user (e.g. pupil dilation and stimulus events against time) could be compared on the same plot to reveal any possible trend or correlation (figure 2.12).



(a) (b)

Figure 2.12: Representatives of one-dimensional visualization: (a) pupil dilation and screen coordinates versus time (URL13) (b) scanpath of different groups of users versus time (URL8).

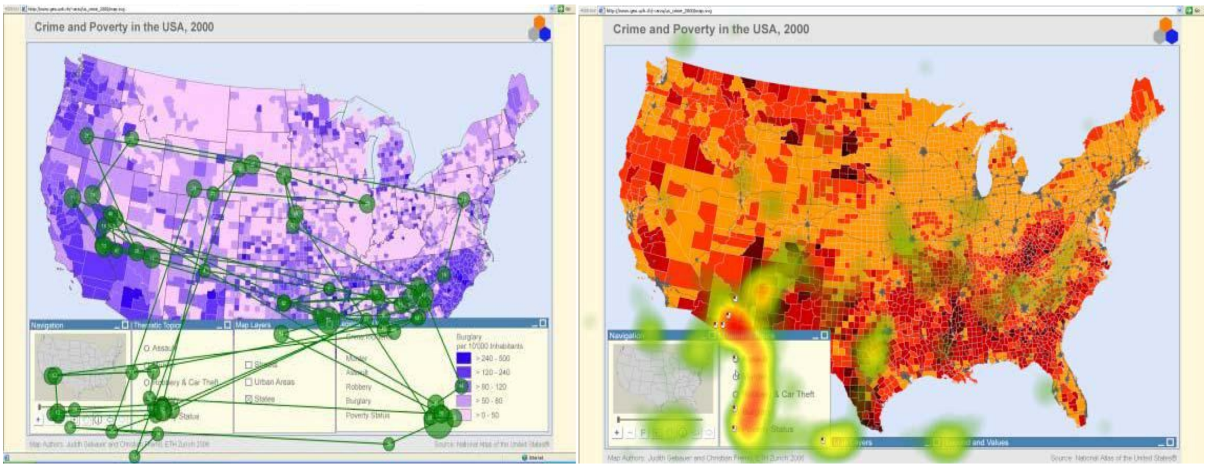
- *Scan plot* is the illustration of plotted eye movement raw data, superimposed over the image of the stimulus. Although scan plots are usually part of the ET visualization tools and used for interpreting the data, since no analysis has taken place yet, they cannot be considered as visual statistical outputs, still they are the simplest visualization technique to implement (figure 2.8-a).

- *Gaze plot* (*scanpath* or *view route*) illustrates a succession of fixation and interconnecting saccades superimposed over the stimulus image as background. The disadvantage of blurred visualization of scan plots, by using a clustering algorithm in gaze plots is overcome. The green circles (figure 2.13-a) mark the fixations in numeric order and their size denotes the duration spent on each circle point. The green lines indicate the saccades. Gaze plots provide a snapshot image of attention during an ET experiment. However, gaze plot is time consuming to analyze and for long sessions it may become cluttered; using a heatmap or dynamic gaze replay (which gives the best the sense of time and order) in this case might be a better choice.

- *Heatmap* (*hot spot map*) is perhaps the most revealing output of an ET study (figure 2.13-b). In contrast to gaze plots which provide eye movement information of an individual user, a heatmap is based on the summary of gaze positions gained from multiple sessions and users. Heatmaps utilize a range of different colors in order to differentiate among the least observed areas to the most. Although there is little difference between the two, heatmaps can be used to plot either fixation count or gaze duration. heatmaps are typically used to give ideas for places to further investigate by for example gaze replays (Pernice and Nielsen, 2009). The transparent superimposed colors of the heatmap facilitate observing the underneath stimulus. However, some tools can just create opaque heatmaps which are not convenient for analyzing the stimulus. Depending on the ET software being used, heatmaps can be applied for the static and also for the dynamic stimulus, which has the same underlying concept, except that the heatmap generating algorithm of dynamic stimulus uses data of multiple recordings from the last few minutes (Špakov, 2008).

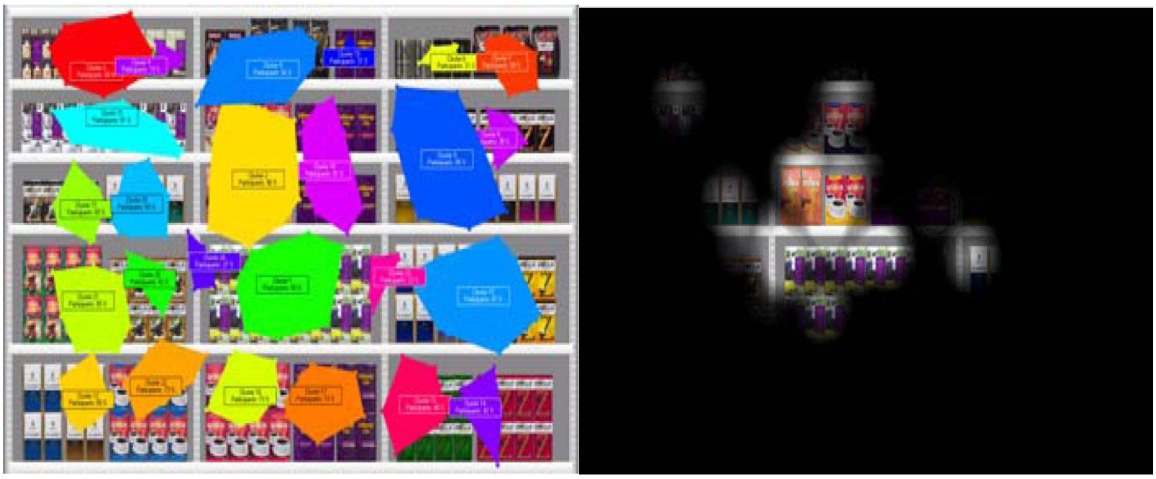
- *Gaze opacity map* (*inverted heatmap* or *fixation map*) has the same underlying concept as a heatmap but with different visualization (figure 2.14-b). Here most parts of the image are blurred and only most heavily observed areas gradually become transparent; the transparency rate is proportional to the nearness to the fixated points (using either linear or Gaussian estimation methods). The disadvantage of this method is that if the total transparency reaches 100% further fixations have no effect on visualization (Špakov, 2008). There is also a new visualization identified as *cluster* which shows areas with a high concentration of fixations as polygons (figure

2.14-a) or opaque circles. To calculate these condensed fixation areas, the software applies some clustering algorithms.



(a) (b)

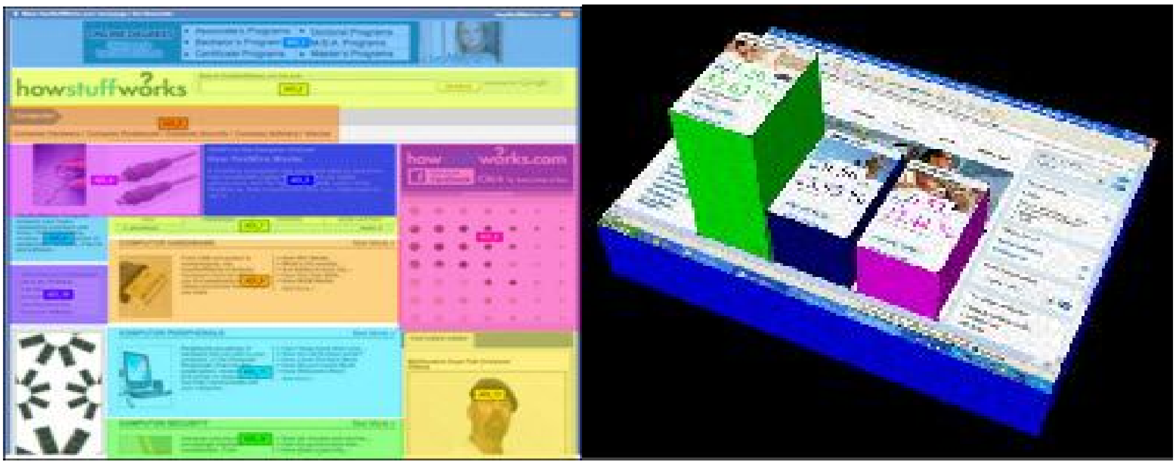
Figure 2.13: Representatives of: (a) a gaze plot, and (b) a heatmap (Çöltekin et al., 2009).



(a) (b)

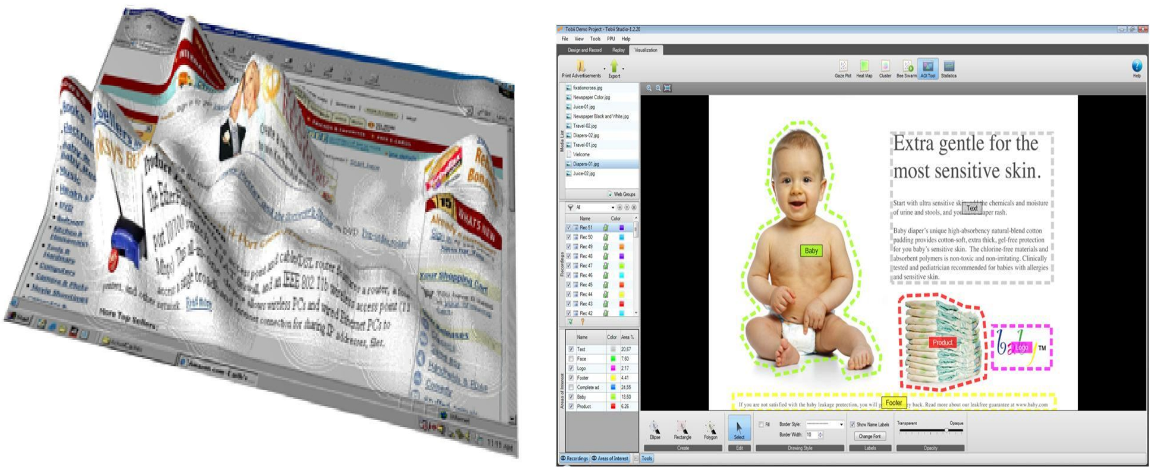
Figure 2.14: Other ET software products: (a) cluster, and (b) gaze opacity map (URL2).

- *Clustering fixations* is a general term for illustrating either the intensity of gaze data in some predefined areas of the stimulus or the highly observed areas by the user in a brief glance. In either case a clustering algorithm is required to group fixations. *Region or Area of interest* (ROI/AOI) is a predefined area on an interface which contains some semantically interrelated information such as legend of a map or some words in a piece of text (figure 2.15-a). Since choosing different areas as ROIs produce different results, these areas should be defined carefully. ROI is typically applied to aggregate and compare quantitative data from large numbers of test participants. The output of the ROI measure could compare: average, minimum, maximum or standard deviation of fixation duration, number of fixation, number of gaze data-point entries, absolute and relative time that users spend in each area and also a list of system or environmental events which happened during observing a ROI. These outputs are usually visualized as charts; although there is a possibility to observe the outputs directly through the image of stimulus. This facility is realized by combining bar charts with the 3D stimulus in which highly observed areas pop up over the rest of the image and the size of raise relates to the quantity of the measure being investigated (figure 2.15-b). ROI may be defined manually (usually as rectangular). However, in some software, highly observed areas could be extracted automatically by some image analysis algorithm (figure 2.16-a). This output has the same underlying logic as a heatmap, but with a *3D representation* in which peaks are the illustration of longer observed areas. The rotated view is for depicting a more efficient perspective. Inaccessible or hidden sections are the disadvantage of this product. Also, in order to measure eye data in a non-rectangular area, there is a possibility of defining ROIs as other shapes in several softwares (figure 2.16-b). And finally, there is a new tool for keeping tracks of ROIs and synchronize them with gaze data in changeable interfaces like web pages as well as for video recordings. This tool allows defining ROIs for each individual page and the gaze data for each recording are automatically adjusted relative to the ROI positions to compensate for scrolling, moving windows and page transitions; in the same way, for video analysis after defining some ROI, the software keeps track of the moving ROI to the extent that it can identify the object (Manhartsberger and Zellhofer, 2005) (Špakov, 2008) (URL2).



(a) (b)

Figure 2.15: Illustration of predefined ROIs of a webpage: (a) defining semantically related ROIs, (b) 3D visualization of comparing three ROI outputs (URL14).



(a) (b)

Figure 2.16: Illustration of two other products: (a) 3D (topographical) view of highly observed areas in a webpage (URL14), (b) defining non-rectangular ROIs in an advertisement (URL2).

Some tools can combine some of the mentioned outputs to produce composite products for special analysis. For instance, it is possible to create the heatmap only inside a defined ROI. Apart from mentioned ET products which are to some extent common among ET softwares, there are some other visualizations that are unique for their software. Table 2.4 presents a summary of ET analyzing software, found in literature, which could provide unique visualization among other tools and give some information about their other visualization techniques.

Table 2.4: Summary of ET analyzing softwares which provide unique visualization (Špakov, 2008)

Tool	Visualisations: 1D	Visualisations: 2D				Visualisations: 3D	Other	Uniqueness
	X/Y/Z vs time	Scan-plot	Fix-Sacc plot	Heat map	Re-play			
Gaze-Tracker (ERICA)			+		+	+	MS Excel graphs, pupil size vs time	Unique visualisations: - 3D bars - AOI tracking in replays One of the first visualisations: - 3D hills and valleys
Eyegaze Trace Suite (LG Technologies)	+		+		+		Pupil size vs time, Replay of mouse movement and keypresses.	Unique visualisations: - Combination of 1D and 2D visualisations
Print.Analyzer POS.Analyzer Web.Analyzer (Media-Analyzer)			+	+	+		Heat map combined with replay, Stimulus + AOIs	Unique visualisations: - Heat map replay of data from multiple trials
BeGaze (SMI)	+	+	+	+	+		Graph of comparative analysis of scan-path of several subjects	Unique visualisations: - AOI observation percentage second-by-second
Clear-View Tobii Studio (Tobii Techno-logies)		+	+	+	+		MS Excel	This tool is included in the list as a very popular and widely used gaze data analysis software.
EyeData-Analyzer (Neuroin-formatics Group, Bielefeld University)			+	+	+		Shadowed, fogged and blurred views. Clustered view. Clustered fixations.	One of the first visualisations: - various "fixation maps" - automatic screen separation into zones and highlighting of the most observed zones
Unnamed (Visual Cognition Lab, Michigan State University)		+	+	+	+		Contour Plot of Fixation Time. Fixation duration and saccade amplitude frequency distributions. The replay using a high-pass filtered background image	Unique visualisations: - "Moving Window" visualisation that replays the recording using high-pass filtered background image
NYAN (Applied Cognitive Research Unit, University of Dresden)			+	+	+	+		This tool is included in the list as an example of software developed in academic institution (now it is distributed by Interactive Minds as a commercial product)

2.6.4. Phase Four

Data interpretation is a crucial stage in an ET study. Without sensible interpretation, an ET experiment produces no result, even if all previous stages are performed correctly. Unfortunately due to reliance of ET objective data on the outcomes of (some of) subjective techniques for interpretation, there is limited validity for the interpretation of ET data. Moreover, there are restricted achievable conclusions about cognitive processing. These factors which are discussed below can lead to presence of different explanations for a specific outcome.

2.6.4.1. Interpreting eye movement data

Relating peoples eye movements to their inner feelings and motivations, also regarded as 'windows of the soul', is not a new discovery. 'Shrewd salespeople, poker players, negotiators, even the family dog, study a person's eyes to discern their emotional state or point-of-regard (Ellis et al., 1998).' The relationship between eye movements and whatever motivated these movements is defined as a hypothesis (Poole and Ball, 2005). Although it does not hold all the times, *eye-mind hypothesis* formulation is the basic principle at the origin of all ET researches. It assumes where a person is looking at, is the indication of what he is thinking about or attending to at that moment.

The last stage of an ET study is data interpretation, which is done through assessment of the data analysis results and making logical comments in order to evaluate or develop hypotheses. Interpretation of ET data is to some extend a subjective process, since it relies on the subjective data of conventional techniques to be confirmed. For instance, after the analysis of the output indicated a large number of fixations on some place of the stimulus, confirmation of some qualitative technique like user complaints through TA, is required to proof that it was due to some unclear icon representation and not because of attractiveness of that part of stimulus. Likewise, during a scene observation, the eye may fixate on a specific element, while the person is actually attending to the whole scene without devoting attention to the fixated point, or maybe he thinks about something totally irrelevant. Due to the limited ability of human beings in judging his unconscious thoughts or recalling them and expressing himself in complex tasks during TA and RTA, such techniques are not always capable reflecting the actual reality. Although combining ET and TA as complementary techniques raise the validity rate of the results up to a high level, there is no guarantee for that. Furthermore, due to different resulting eye patterns, the interpretation sometimes may be different if user is TA versus not TA.

Another issue in interpreting ET data is how to relate this data to cognitive processes and to assess the influences of the context of the user interface that is being evaluated (Manhartsberger and Zellhofer, 2005). There is a distinction between top-down (task oriented) and bottom-up (behavioral inferences) methods in traditional information processing frameworks. The interface exerts a bottom-up impact on user's eye motion behavior that is triggered by salience of the interface itself. Task requirements and user memory about 'what is where' from previous experiences which represent the top-down factor is also having an influence on eye movement behavior (Henderson, 2003). Top-down factor itself could be based on either cognitive theory or design hypothesis (Goldberg et al., 2002). While it is achievable to model the influence of the bottom-up processing (Turano et al., 2003), obtaining data through ET and other methods to model top-down factors is problematic. Although it may seem logical to infer cognitive processes from eye movement patterns, there is not always a strong hypothesis or theory to guide the analysis (Jacob and Karn, 2003). However, it might be possible for simple clearly defined tasks, to model such top-down cognitive processes (Mayr et al., 2009).

Looking at previous ET experiments in usability evaluation shows that both approaches are applied. The interaction between these two processes results in *information scent* which is perceived when the proximal cues (bottom-up) are evaluated relative to the current goals (top-down) and if ET is to become a routine usability technique both styles of ET studies must be adopted (Goldberg, 2003) (Zambarbieri, 2005).

2.7. Advantage, disadvantages and pitfalls of ET technique

Like any other usability technique, ET has its weak and strong points and deciding on its functionality as a whole is not quite simple. ET can reveal valuable information about hidden cognitive processes of human behaviors for recognizing where, when and what exactly users have looked at, which may not be reached by conventional techniques (e.g. the reason that users failed to spot some element might be influence of another distracting element or just its inappropriate position or design; each case should be treated differently). ET can also tell us whether users are looking at the screen or not, differentiate reading from scanning, ensure the design quality by revealing relative intensity of a user's attention to various parts of a prototype (Namahn, 2001), reveal user's personal interests by detecting the mismatch between the expectation of designer and the behavior of user (Meng, 2004), help to explain individual differences (regarding their aptitude, expertise and even pathology), determine whether a user is searching for a specific item or just browsing by detecting increased pupil diameter during browsing, compare users' overall scan patterns, demonstrate scanning efficiency, support other types of data, help to discriminate dead time, understand expert performance for training others, help to sell usability testing, assess users' decision making processes and provide domain specific benefits (e.g. web pages, cockpits and text design) (Namahn, 2001). ET is also useful for testing hypotheses about design principles such as contrast, repetition, and proximity (Cooke, 2005). Contrary to conventional techniques which are subjective and qualitative and reflects users' conscious thoughts and feelings or the observers' impressions, ET's objective and quantitative data, provides a more scientific basis for a usability testing. Considering these benefits, some researchers concluded ET can enhance usability testing by combining it with other conventional techniques such as interview and questionnaire (Çöltekin et al., 2009).

The technique also has its limitations. First, it requires an ET system which is not always affordable; still the price should be reduced gradually as technology evolves. Different applications require different amounts of system sensitivity; unfortunately there is always a trade-off between spatio-temporal accuracy versus cost and ease of use. Another issue is that it requires experts to operate the system properly and to interpret the results. ET results would be applicable only if the technique is applied correctly, which often is not the case due to the lack of knowledge or experience (Pernice and Nielsen, 2009).

Some other issues include: integrating ET data with data from conventional techniques for obtaining optimum results (Jacob and Karn, 2003) as well as standardizing which metrics are used and how they are defined and interpreted in order to be able to compare different studies on an even footing. Technical issues like the minimum time for detecting a fixation (also limits of accuracy, resolution as well as defining areas of interest or ROIs) can affect interpretation of cognitive processing (Poole and Ball, 2005). ROI should be defined in a way that to be able to capture all relevant eye movements. Although in most cases the eye-mind hypothesis holds, it is not always true, meaning that ET cannot determine why users are looking at something or even prove that users did not see something since users can acquire information through peripheral vision (Bojko and Schumacher, 2008) ('Peripheral vision is the ability of the eye in order to monitors a visual field of about 200º, while receiving detailed information from only 2º which is called the fovea.') (Richardson and Spivey, 2004). Moreover, eye movements involve some uncontrolled behaviour and unfortunately there is not much evidence of gaze behaviour to distinguish between a conscious or deliberate fixation and an unintentional one (Posner, 1980). Some systems are accurate to within one degree of the point-of-regard. The problem is that attention can be directed, without any eye movement, up to one degree away from the measured point-of-regard. Also, handling excessive volumes of ET data for complex tasks is not so easy and the software for data collection, clustering and analysis are not streamlined to operate together automatically (Poole and Ball, 2005). Although some initial efforts (Bates et al., 2007) towards standardizing raw data formats and their accessing tools have been started, there is still a lack of interoperability among different ET tools and manufacturers and most of ET software are highly dependent on the eye tracker's type for their raw data format and data gathering methods (Špakov, 2008). Eye tracker's output raw data are seldom pixel perfect, and due to the lack of data validity sensors, before filtering they need some reviewing and correcting for two types of errors: absolute drift in which all the points have drifted off together versus relative warp which are harder to correct (URL_3); it is not always the case that the system exert this corrections automatically, when recalibration can prove costly, manual or automatic methods are required to compensate for errors. In order to reduce intrusiveness, designing head mounted systems should be improved to make lighter systems; although IR (infrared light) is not sensed by the eye, there is the problem of visible light reflection through the mirror part of the glasses mounted optics in outdoor applications, which is bothersome. In the same way, fixed systems should allow more head movement without loosing calibration (Poole and Ball, 2005). Calibration opens another category of problems of ET. Some people cannot be eye tracked properly; sometimes the study requires calibrating of participants with low attention spans e.g. babies. Sometimes the pupil may not reflect enough light, it may be too large, too small or it may be occluded by long eye lashes or eye lids; the person may have a wandering eye, lazy eye, low pupil contrast, glasses or contact lenses. A person’s eyes may dry out during the test. A user may be eye tracked one day and not the next. Also, maybe the ET, half way through a test degrades to the point that the collected data cannot be used (Namahn, 2001).

Using the ET technique also involves some pitfalls regarding the data collection and analysis tools like system errors, which causes no data recording (usually loosing 1 out of 10 tasks per session), hard crashes which takes an inordinate amount of time to reboot, the system might be slow to load or save files and can run just one process at a time, the systems are bad at giving feedback information to indicate any sort of progress or errors, which is important when dealing with large files (Pernice and Nielsen, 2009).

Also for the analysis, one of the pitfalls of available tool is the ambiguity and lack of open processing method regarding their built-in clustering algorithms. Another new challenging issue regarding data analysis is the mapping of fixation points to visual stimuli or gaze data labelling for different dynamic stimuli including video captures of physical environment or the monitor screen. This process, which requires a robust object extraction (*objects are graphical units that have some data-driven attribute*) image processing algorithm to bind gaze data to the moving object specified, is still in early stages (Špakov, 2008).

2.8. Conclusion

Eye movements are considered as the result of both properties of the visual world and a person’s mind processes. Concerning the temporal dynamics and psychological processes, they can provide a rich data source that can be led up to the response. These findings make ET highly applicable in a variety of task domains. Although still in their infancy, this information also can be fed back into devices in real time, in order to issue instructions or adapt computational processes. Likewise, areas like personal computing, the automotive industry, medical research, and education will soon be utilizing ET in ways never thought possible.

In usability studies, ET data allows for detailed measurements of how a user is interacting with a system. Considering different opinions upon the of effectiveness and efficiency of ET in usability researches, there is almost a mutual agreement on ET enhancing traditional techniques by giving new insights into those issues that are already known. In cases where users are not proficient enough at verbalizing or reflecting on their own behavior, without ET, design recommendations have to be implemented by trial and error, adding to development time and cost (Cooke, 2005).

On the other hand, considering the cost-benefit perspective, many researchers regard ET as a valuable component of the usability toolkit, but just for companies with a highly evolved usability culture and large budgets that have already covered all the basic issues with cheaper non-ET techniques and a minor improvement in products’ usability for them could save a lot of money. For others, instead of ET quantitative data, they prefer to rely on traditional qualitative analysis like watching user’s gaze or TA listening and ones insight (Pernice and Nielsen, 2009) (Namahn, 2001).

Among all these judgments, in order to be able to achieve a more realistic view of ET situation and its applicability as a user research technique, especially in geo-domains, it will be tried in the next chapter to experience this technique as two case studies and observe the potentials and limitations in practice.

3. Test design for usability assessment of ET in the geo-domain

3.1. Overview

The previous chapters were devoted to describing common characteristics, capabilities and limitations of a user research technique which is ET. In this chapter we are going to develop two case studies in order to investigate the usability of ET as a user research technique in geo-information processing and dissemination in practice.

The first stage of usability testing is to design and plan a valid test, such that it can fulfill the objectives it is initially designed for and leads to desirable results. If the test is not appropriately planned, other stages will fail as well. The present chapter discusses the design stage of two planned case studies with the two main types of eye trackers (fixed and mobile) based on the obtained knowledge from the literature review regarding the characteristics of the ET technique and other sources of information (e.g. consulting ET companies), in order to assess the usability of ET for a geo-application in practice. The first case study tries to investigate the usability of mobile ET in obtaining extra user information for pedestrian navigation systems, compared to other user research techniques like audio-video recording of the user. Likewise, the second case study looks into the usability of a fixed ET system in finding extra user information for a geo-web application.

3.2. A case study with a mobile eye tracker

3.2.1. Introduction

This chapter starts with depicting the motivations and the aims of designing the case study. Then, the pre-test procedure in which the equipments are applied and examined in order to do the actual test is explained. The pre-test experiments lead us to find some limitations regarding the equipments' specifications and regarding what is expected from the equipment in the already defined case study. Because of the mentioned limitations, the case study is then adjusted and redefined in details in the next sections. This includes defining the aim of the new test and the questions that the test is supposed to answer, variables of the usability testing, participants, environment and study area, scenario and techniques.

3.2.2. Mobile-ET user research methodology in the requirement analysis phase of pedestrian navigation systems

The purpose of this case study is to provide new information about the possibility of applying the mobile ET technique in the geo-domain. In the present case study, a user test is going to be executed for a mobile geo-application that is later going to be designed. The design of this system is based on personal geo-identification, which is the relationship between reality (environment), the representation of the reality (screen of the device), and the mental maps of the users (Delikostidis and van Elzakker, 2009). The question is to what extent the mobile ET method and its resulting data are usable to create information for the design and implementation of this application.

Usability testing has proved to be an important part of the design and implementation of any prototype such as the mobile geo-application. Delikostidis (2007) provides a new field-based usability testing methodology for mobile geo-applications and pedestrian navigation systems. He utilizes an advanced technical solution (which is a prototype for audio-video recording of the user via imbedded cameras, etc. in a hat and a bagpack which the user wears) to support his methodology, which reduces the use of human resources and time needed for user data collection and allows for better analysis of the results.

He compares the use of this method with the use of other combinations of user techniques, in a field-based usability testing. ET was not included in this study. In his comparison, he mainly investigates the usability of zooming, panning, rotation and orientation functions and the quality of map scaling of mobile geo-applications. The outcome of his research supports the use of the new methodology as the most informative and resultful method. Delikostidis and van Elzakker (2009) in a field-based experiment, investigate the interactions of users in unfamiliar areas with the environment, their mental maps and the interface of the geo-mobile applications via the above mentioned technical solution. The results of the field-based usability testing show the importance of GPS-independent automatic map orientation and of particular landmark (LM) types for geo-identification and navigation.

Considering these studies and in order to assess the usability of mobile ET in providing extra user information in a scenario compatible to the aforementioned, we decided to use mobile ET (alone / along with other techniques of collecting user data). In order to do this, in the present case study a PDA (i-mate ultimate 9502) running iGO 2006 (a pedestrian and vehicle navigation software) is applied in a field-based usability testing consisting of a navigation task from one point of interest to another. This navigation task is executed by 2 different groups of users. Each group makes use of different methods for data collection during the navigation. These methods include applying the ET technique alone and the ET technique along with TA (see section 3.2.1.3). The outputs of the two groups will be analyzed and compared to each other in order to find the most informative and usable method for a case like this, but-most of all- to be able to draw conclusions about the usability of the ET technique. The outcomes may help in choosing proper techniques to be used in the related PhD research that will lead to developing a similar (like iGO 2006), but improved pedestrian navigation prototype, more suitable to users' needs. This means, the present case study just evaluates the use of different methods (eespecially ET) for usability testing, in providing informative data in the requirement analysis stage of the UCD process of such an application; but the actual design is not part of this thesis.

To summarize the above arguments, the ideas for designing this case study are threefold:

- First, the main purpose of this case study is to investigate the usability of the mobile ET technique in the geo-domain (this main purpose also applies to the second study which investigates the usability of the fixed ET systems). For this purpose, ET (alone/combined with other user research techniques) is applied in a field-based user testing in order to find the most informative approach in providing user data for a mobile geo-application.

- Second, the test is designed in a way to cover a new application of the ET technique in the geo-domain. Moreover, this case study fits to a broader project which contributes to the improvement of pedestrian navigation and geo-identification (Van Elzakker et al., 2008), as indicated above.

- The third factor is to investigate the presently available facility in ITC (ASL Mobile Eye tracker) which we can apply, and to take into account the capabilities and limitations of the system in designing the test.

Considering the third factor, we should know more about the equipments we are going to use for the test, which in this case are the ASL Mobile Eye tracker as well as the PDA. In pre-test experiments, in which we tried to familiarize our selves with systems' operation, we ran into some problems which made us to change the direction of the case study a little, such that we can still focus on assessing the usability of the mobile ET for a geo-application. These problems along with the specification of the related devices are explained in the next sections.

3.2.2.1. Test equipments

Selecting proper equipments in a usability testing is an important factor that influences the possibility to collect the optimum possible amount of relevant user data with the highest quality. Selection of proper equipments relates to the selection of correct methods and equipment for the test and also the type of the product under investigation. Finding out about the right equipments and techniques is one of basic purposes of this case study. To perform this case study we ran into some problems in applying the equipments including eye tracker and PDA. The issues we met regarding the eye tracker include calibration, scene camera issues and the lack of required equipment (Laptop instead of PC). We tried to overcome these problems through pre-test experiments, literature, and communication with the manufacturer (ASL) through email and an on-site training session. I should mention here that ASL's staff were very accessible and willing to help us to solve the issues. Problems regarding the PDA include small screen size and the environmental light reflection from the screen. Obviously, the encountered problems regarding the eye tracker and the PDA are just related to this case study; they may not be the same for other applications. All aforementioned issues will be discussed below.

ASL Mobile Eye

Before giving details of encountered issues, we first explain characteristic of ASL Mobile Eye tracker.

System specification

ASL Mobile Eye is the first tetherless and compact head mounted eye tracker (figure 3.1), manufactured by Applied Science Laboratories, a company which, having over 30 years experience is the first ET manufacturer to provide head mounted eye trackers. ASL offered products are based on either VOG or IROG techniques (see section 2.6.1.3). VOG based products include both remote and head mounted eye trackers (see section 2.6.1.1). There are two presented designs for the optics in head mounted systems; the optics are integrated into a helmet or they are goggle mounted (as in Mobile Eye). ASL systems mostly apply the bright pupil method which is supposed to work better than the dark pupil (see section 2.6.1.3). The bright pupil method is more adjustable to different types of eyes (by creating less noise and fewer track losses) and can also be used in a darker environment. However, it might be more sensitive to the ambient illumination of daylight. The Mobile eye tracker measures user's eye movement with respect to the head. This measurement is displayed as a superimposed cursor over the image of the scene camera (figure 2.9-b). Produced data include *x* and *y* eye position coordinates, time, and pupil diameter. The ASL Mobile Eye is a bright pupil, VOG based system designed to work for different applications and anywhere (indoors and outdoors) during the performance of natural tasks whereby the user should not be restricted in head or body movement, he should be able to wear head mounted optics.

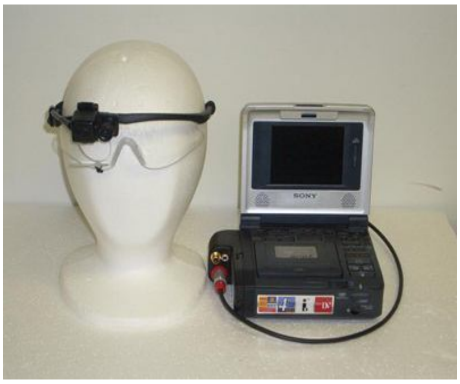
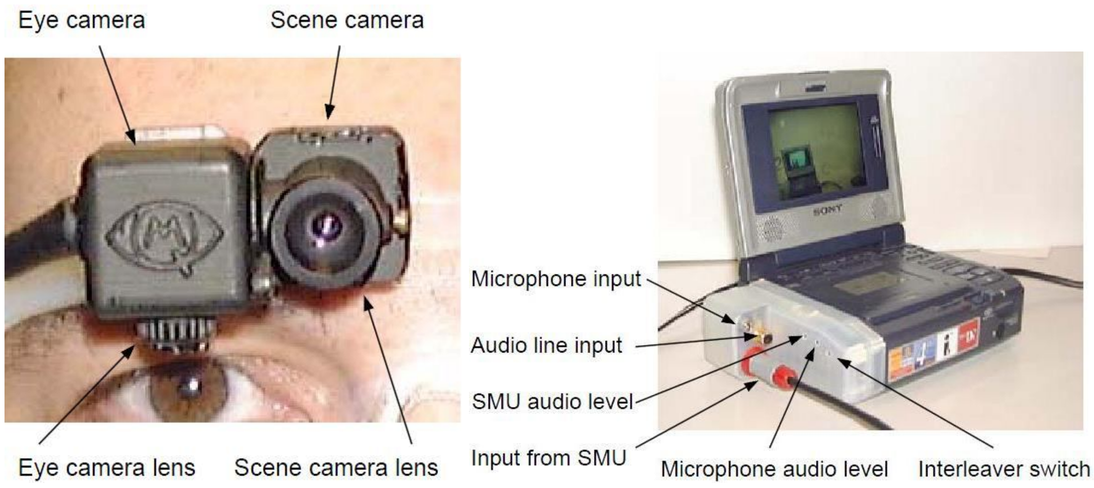


Figure 3.1: ASL Mobile Eye tracker configuration.

System components consist of: *head mounted optics* including *glasses*, *optics module*, *monocle* and *colour scene camera assembly*. The optics are mounted on an adjustable headband above the right eye, which is usually the dominant eye (figure 3.2-a). An adjustable monocle or *hot mirror* (which allows reducing spectacle reflections and decreasing challenges with corrective lenses) is attached to a head band with a flexible boom arm. The head mounted part is lightweight (76g) and the recording device is small enough to be worn on a hip pack (figure 3.2-b). Other components are: modified portable VCR recorder battery operated (130 mins duration), Eye Vision software (which is used for doing the calibration, adjustable to individual users' eyes), PC Pentium4, 2.8GHz processor or laptop w/3GHz processor. The eye image and scene image are interleaved and saved on a VCR tape (60 mins duration); the video is then transferred to a laptop with the installed Eye Vision software, and the calibration is completed via this software.

- Calibration of ASL Mobile eye: the calibration process involves two main steps: first the scene camera axis is adjusted to the eye sight axis. To do this the user is asked to wear the eye tracker and to look straight towards a point. Then, the hot mirror and the cameras are adjusted while participant is looking straight and keeping his head in a stationary mode. The adjustment can be checked either by looking at the VCR screen (in the field), or by looking at the computer screen which is running Eye Vision software; it includes putting the 3 infrared dots emitted from the eye camera, inside the pupil manually (by moving the hot mirror and cameras). The second step of calibration process includes adjusting the gaze point location to its location in real world. To do that, we start recording. Then 'the spot and pupil calibration' is done in software while the user is still looking straight. Then the participant is asked to look at some specific points (at almost the same required distance for the test) for a few seconds while he is still looking straight and keeping his head in a stationary mode. The point at which the participant is looking is shown simultaneously on the screen, and the experimenter should click on each point while the participant is looking at them. After this, calibration is done and the participant can move. To check calibration accuracy, the participant is asked to look at some other points. The procedure requires the attention and patience of participant and experimenter and if calibration is NOT acceptable, the whole process should be repeated. If calibration is acceptable, the participant is asked not to touch the glasses any more throughout the test, and he will start the test. Other mobile systems also apply roughly the same calibration process.

The final output of this stage is a scene video with a calibrated cursor overlay. If all the stages of the calibration procedure went well, this cursor should now point to the actual viewing point of the user. For real-time tracking the portable VCR can be connected directly to a laptop. Multiple or parallel studies is possible by using additional stand-alone optics and VCR recorders (this is for the applications which require to assess reactions of different people on a same theme and at the same time. In marketing for instance, in order to compare 2 people's eye movements while they are looking for the same stuff on the shelves).



(a) (b)

Figure 3.2: Different components of ASL Mobile Eye: (a) The light weighted head mounted optics are attached to a head band above the right eye, (b) A rather small recording device which the user should carry in a waist bag.

System's sampling rate is 30Hz, the measurement principle is pupil-corneal reflection (see section 2.6.1.3) with custom outdoor enhancements, the system accuracy is 0.5 degrees visual angle, resolution is 0.1 degree visual angle, and visual range is 50 degrees horizontal and 40 degrees vertical.

Calibration issues

The main encountered problem in using ASL Mobile Eye was the calibration issue. The system is very sensitive regarding the calibration procedure. Although every time we tried to follow the same calibrating procedure, we could not always get the same results. Sometimes the calibration outcomes in two successive trials for the same person were quite different, and occasionally it was totally unacceptable. Regardless of these occasional happenings, we found out during the experiments that the system provides acceptable calibration accuracy for close range objects as well as for distant objects separately. For our application, however, the user is required to look alternatively at the scene (real environment) and at the PDA screen. So, we needed a system which can provide good calibration for two different foci during the same test. We found out that unfortunately, the ASL Mobile Eye could not provide the required accuracy for this particular case study regarding calibration for two very different foci distances. This means whenever we tried to do the calibration for a long distance, we lost the accuracy for inspecting near objects and vice versa. Through the communication with the ASL company and testing of the system, we found out that the different accuracy for distant and close surfaces is due to a *parallax error*. The ASL Mobile Eye system utilizes the head mounted scene camera for capturing the scene image. Since the positions of the eye and scene camera are separated, the perspective of the scene image is not exactly the same as the perspective of the subject's eye and they have different *'paths'* to the calibration surface (we can imagine a small inclination angle between the sightlines of the eye and the scene camera). These paths converge on the calibration surface. But if the distance changes (e.g. when distance to an object is very different from the calibration distance), we observe a parallax effect or offset between the scene image and the eye direction. The only way to overcome this issue in such a system would be for the scene camera to be embedded in the eye which is impossible. To overcome this issue we had to follow one of the following directions:

- One solution could be to perform two separate calibrations for the same subject: one close and one further away. These calibrations should be created and saved under different names. Later for the analysis then, we should load the close calibration while the subject is looking at the PDA, and load the far away subject profile while he is looking at a distance. Since the execution of this solution for the data analysis is too complicated, and also due to the problem of close range calibration (which is explained in this section) we did not use this method.

- Although the best accuracy is always found at the same distance as the original calibration distance, it is possible to get reasonable accuracy for both close range and distant objects by doing the calibration at a middle distance between the two, and compromise on the parallax effect. The recommended distance for such applications like ours is of around 3 meters. Moreover, increasing the number of calibration points (e.g. up to 9) is an accuracy increasing factor. Another important factor in getting a good compromised calibration is that the target points should occupy around 60% of the screen when calibrating, and also the target points need to be at least 10 degrees visual angle apart. Although we tried several times to get the optimum accuracy in this method by changing calibration distance, visual angle between targets, and changing related parameters' thresholds, still the parallax for the close range objects was too much to be used to read a PDA.

- Another way to solve this problem is to perform calibration at the distance of primary interest. To choose the primary interest for this application we noticed that: first, the map objects on PDA screen have smaller sizes than the real world's objects. They also occupy smaller areas in the scene camera's field of view than the surrounding objects, when the user is looking at them during the test. So, close range objects require more accuracy regarding the eye-cursor location. A little shift in the position of the eye-cursor over the PDA screen makes a big difference compared to real world's objects (e.g. if the distance between two specific LMs in the real world is 10 meters, the same distance on a mobile map is 5 mm). Secondly according to what ASL declared, after we do the calibration for a close distance, a small change in the relative distance between the eye and the calibrated surface will cause a more significant amount of shift, compared to when we do the calibration for a longer distance (meaning that parallax error is conversely proportional to the distance of the gaze point from the eye). This difference is regardless of the device type and it is due to a geometrical reason. Due to these reasons we decided to do the calibration for close range distance and to ignore losing the accuracy for longer distances. Although calibrating to close range distance suits best for our application, during pre-test experiments we realized that we cannot get enough accuracy even for close range distances to distinguish at which features the user is looking on the PDA map at any moment during the test, e.g. when the user naturally turns his head or moves his hand a little as he walks. In other words (as it is explained it section 3.2.1.2), the application of ASL Mobile Eye for the usability assessment of PDA systems in a stationary mode is possible. But a walking user creates extra movements which changes the relative distance (between the eye and the PDA) and angle of original calibration a lot. This creates a random shift in position of the eye-cursor on the PDA during the test.

Because of the already mentioned problems, and due to the fact that we still intend to investigate the usability of the mobile ET technique as part of this research for a geo-application, we decided to change the scenario a little (by eliminating the ET data processing for the PDA and concentrating more on real world objects) and to make the primary interest in distant objects. Based on our experiments and on communication to ASL we set the primary interest calibration distance at 3 meters.

Some other findings throughout experimenting with the system, to optimize ASL Mobile Eye calibration results include:

- During calibration, the eye should be constant (rest at the target without vibrating), while concentrating on each target. In order to raise the quality, it is better to pause for some seconds on each target. It is also important, as the final stage of calibration, to make sure that the software accepted the selected target. Sometimes, the software did not accept a target point (it happened randomly). Getting a green cross after the mouse click on the target, means that target point is accepted by the software. Also, reducing the size of the cross hair in the configuration (to around 15 pixels) helps in increasing the accuracy.

- Depending on the expected accuracy, the number of target points could vary (e.g. 4 to 9). It is also important to find a compromised distance between targets. Too close target points reduce the accuracy; likewise choosing spots far from each other may cause some of them not to be visible from the scene camera. Depending on the primary calibration distance, the appropriate distance between targets can be obtained by some trial and error.

- One more distinctive factor to raise the calibration accuracy for close range applications is that if the user is going to look at an object which is not flat (in other words, there is a rather constant angle between point-of-gaze line and the surface under inspection during the test), then the calibration should be done on the same angled surface. For instance, if the participant usually holds a newspaper at a specific angle to read, then the best accuracy for inspecting his eye movements during reading is obtained by doing a calibration while he is holding and looking at a surface similar to what he used to do.

- It is also critical for the eye to be in focus when calibrating. In order to set the eye to be in focus, the three infrared dots should be placed manually in the center of the pupil; good focus on the eye can be obtained by moving the *hot mirror* or the *focus lens* on the eye camera.

- In order to have a stable focus state during calibration, it is important that the camera assembly is not loose in the dovetail and does not move up or down during calibration. It is also critical that the whole head mounted optics does not move on the head during the calibration or the test.

- It is important that the pupil is being recognised correctly during calibration. The loss of pupil recognition during calibration can be recognized if the magenta circle (which shows the pupil diameter) looks huge or if it is frequently lost. The loss of pupil recognition will result in an offset error. To solve this problem, in pupil setting the threshold should be adjusted till it shows a yellow stable cross inside the pupil circle. Moreover, dim environmental light or physiological pupil structure may cause a large pupil size which also leads to loss of pupil recognition.

Regarding the parallax error issue, some other companies which we contacted claimed that they solved the issue of parallax error in their system. *SMI* (URL_8) claimed that although their system is in many ways a better solution compared to the ASL Mobile Eye, concerning the issue of two different foci they have the same limitation. *ISCAN (URL_9)* claimed they overcame this problem with an automatic parallax adjustment feature in their *OmniView* Mobile system (mentioned price US$39,905). Also, the *Arrington Research* company (URL_10) proposed that their binocular *EyeFrame* scene camera systems include a feature for compensating for parallax error where the operating distances vary from the initial calibration distance; they recommended item *BS07* (which is only available as a PCI card system) plus a wireless option and mentioned that the price for the whole collection varies depending on the operating distance from user to the computer and also whether it will be line of sight or through barriers. *SR Research* (URL_11) is another company that claimed their head mounted *EyeLink II* does correct for parallax error and provides much better data quality compare to ASL Mobile Eye. However their system is not applicable for our test, since it is tethered and there are cables coming from the eye tracker's head gear to a data collection computer which allows user to walk up to 5 meter from the computer. They claimed that for further movement all the equipments could be placed on a utility cart with a power source (mentioned price € 28,800). Looking at literature, web pages and contacting different companies showed that among some companies like ASL, ISCAN, and, SMI we probably cannot choose a clear winner. However, a more accurate judgment among such companies requires some trial and error with different systems to figure out which is the best. Such kinds of comparisons require a lot of time and sources and have not been done yet. However, we can find a few instances in literature in which 2 or 3 different eye trackers were compared and used in practice, e.g. (Nevalainen and Sajaniemi, 2004). Also, we can find rather complete comparisons of different eye trackers' specifications regarding the software, hardware, manufacturer, etc in literature (Merchant, 2001) (Špakov, 2008). Choosing a system also depends on the specific application it is used for. ASL is one of the famous companies for mobile ET systems. Based on communications with different companies and other sources of information, the researcher believes that for this particular application, all other comparable eye trackers to the Mobile Eye probably have the same limitation and using ET for pedestrian navigation (studying PDA and environment while walking) with available eye trackers is not possible yet.

In order to confirm pre-test experiments and eliminate remaining uncertainties regarding calibration issues, the researcher attended a training session at ASL UK branch office. It was discovered during the session that the system we use, is around 3 years old. The new Mobile Eye systems are provided with a laptop and the head mounted part in new systems is improved. Regarding the software, there is an added software component, namely *shift calibration* for improving the parallax error. However, the trial with the new system could not provide sufficient accuracy by reducing parallax error for our application. During a discussion about other companies' feedbacks regarding the parallax error issue, ASL clarified that parallax error is a hardware problem which is resulting from the existing distance between scene camera and the eye. What other companies declare as a solution for that is a software solution like the one provided in new versions of ASL software. Unfortunately, it was not possible to go for a deeper exploration of other systems, due to the time and budget limitation in this research.

Scene camera issues

Apart from calibration, there were a few problems regarding the ASL Mobile Eye scene camera:

- The scene camera is auto-adjusted for different lighting conditions. Moreover, it is possible after recording to adjust both the eye and scene cameras for different lighting conditions in the software. Still the observable contrast was low, which occasionally made the outdoor recordings seem dark; this happened eespecially half way the test, when the batteries were no longer fully charged. Because of this, we tried to carry out the test in brighter hours of the day (we finished the sessions every day - in February - before 5 pm).

- Although the visual range of the camera (50 degrees horizontal, 40 degrees vertical) is better than of some other mobile eye trackers (e.g. SMI iView: 30 degrees horizontal, 25 degrees vertical or ISCAN ETL500: 25 degrees horizontal, 25 degrees vertical), it is not satisfactory for our application. A navigation task with a PDA requires the participant to look at the PDA and also to look ahead. During pre-test trials, we noticed that the scene camera's viewing range for a walking user who holds the PDA naturally is not enough. Although it is possible to adjust the scene camera at a compromised angle (between straight looking and looking at PDA), the user should always remember to hold the PDA higher than naturally is the case. Otherwise the PDA would be out of field of view. We noticed in practice that if the user changes a little his viewing angle by turning his head or moving his hand position (to the right, left or downwards), the PDA would be out of the field of view.

Other issues

The particular ASL Mobile Eye system we used for the test includes a PC computer instead of a portable one (laptop). During pre-test experiments we frequently needed to recalibrate. As mentioned earlier, although every time we followed the same calibrating procedure, some times the calibration was not acceptable. Since the experiment is going to be performed in the field which is far from the lab, the calibration process has to be partly (the first part) done in the field. In case there is a calibration problem, it would not be possible to check it before starting the test. This causes a big problem since the whole test should be repeated due to an unacceptable calibration.

PDA and the navigation software

In the following sections first the specifications of the PDA and the navigation software which we used for this test, and then the encountered problems using them will be discussed.

System specification

The system used for this research is *i-mate Ultimate 9502* which is a PDA Phone with GPS. The navigation software is *IGO my way 2006 plus*. The reasons for using them are their availability to the researcher and the simple functioning of zoom, pan and rotation which are the only required and applied functions for this test.

Small screen size issue

The screen of a PDA is usually small. Consequently the applied map on it, at a normal zoom has a small scale, and the user who wants to use it, normally keeps the PDA at a close distance of around 30 cm from the eye. Although ASL Mobile Eye and some other eye trackers have been applied for the inspection of hand held devices before, all of them used the application in a stationed state, while the user position is constant and the hand which holds the application is stable (usually resting on a table). As far as we know, ET has not been used for a pedestrian navigation task with a geo-mobile application before. As mentioned earlier, the parallax error is conversely proportional to the distance of gaze point from the eye. The error gets more significant as the object gets closer to the eye. However, using a PDA in a stationary mode and at the same distance and angle from the eye (as during the original calibration or primary interest calibration) increases tracking accuracy. If, however, the user is walking along a path (and TA), it is impossible for him to always manage to hold the PDA at a rather stationary mode and this decreases the accuracy.

In contacting ASL, we found out there are three additional lenses for the scene camera of Mobile Eye which can help in improving the accuracy of inspection of close range distance objects. However, these lenses are just designed for close range distance applications and the draw back of using them is a narrower field of view. It means the lens which provides the most accurate position for nearby gaze points, has the narrowest field of view. This solution was not applicable for this case study, since we already had a problem, even with the widest possible camera view field (figure 3.3). For this application, the user is required to look alternatively at scene and at the PDA and using a lens increases the chances of the PDA falling out of the camera view field.

Figure 3.3: Illustration of the usual use of a mobile eye tracker (SMI iView) for the PDA inspection, while the user is in a stationary state. The user's field of view can be observed on line, by the related software. It can be noticed that the field of view (even without any lens) is small and if the user moves his hand a little, the PDA falls out of the camera view field (URL_8).

Environmental light reflection issue

Apart from the size issue of the PDA, there is the problem of environmental reflection of light from the PDA screen which can make it hard to see the screen clearly (figure 3.4). The reflection can be reduced by using anti glare stickers which are available for use on small screens including PDA's. A better solution to improve this issue is to integrate and synchronize the video output of the eye tracker with a separate recorded screen-logging of the PDA. However, this would be just a partial solution, since by screen-logging it can be seen what information on the whole screen the user is looking at, but the eye-cursor is not visible.

Figure 3.4: Illustration of environmental light reflection from the screen of a mobile application.

The aforementioned issues, and mainly the unsolved calibration problem for close range distances make analysis of the gaze-data over the PDA (like the usability of the rotation, zooming and panning buttons, their iconic representations, iconic representations of LMs, tracing LMs back on the map, using navigation arrow, etc.) difficult. The conclusion is that we cannot use ASL Mobile Eye (or maybe other mobile eye trackers) for the implementation of a scenario as originally intended (see section 3.2.2). So, we decided to adjust the case study and to change the scenario a little by setting the main focus of the calibration on real world objects, instead of the objects on the (carto-) graphic interface of the mobile geo-application. The following sections describe the adjusted case study for the usability assessment of ASL Mobile Eye tracker using a mobile geo-application, in details.

3.2.3. The adjusted case study

In order to design a usability testing properly in general, the *aim* of the test, the *designing stage* of the product onto which the test is carried out and the concrete *questions* that the test is going to answer should be specified. These questions then would be translated into *tasks* which user should carry out during the test and finally the results of the carried out tasks would be used for the development of the research prototype (although implementation is not a part of the present case study). Other important factors in designing a test include proper selection of *variables* of the usability testing and the *techniques* for measuring these variables, *interface* and the *functions* to be investigated, *equipments* for data collection, *test procedure*, test *environment* and test *participants*. These parameters can be modified according to different circumstances of different tests.

One of the mentioned distinctive factors in designing a case study is the stage of UCD at which the product is going to be evaluated. As explained in chapter two, the first phase of the UCD is the requirement analysis, which is an essential stage of any user centred design project. Requirement analysis includes a thorough investigation into the needs that a system has to fulfil (Lamsweerde, 2000) by assessment of users, involved tasks and the context of use. The aim of the requirement analysis phase is to identify the user's basic requirements, satisfaction level and encountered problems in using a product. This can be an existing product or it can be achieved by adding extra functionalities to a present system or making some improvements on it, and then doing the usability testing to obtain required information for designing the actual product. The result of assessments in this phase later leads to initial design of the product in the product design solutions phase. As already mentioned, this case study is going to assess the usability of mobile ET for the pedestrian mobile geo-applications in first phase of the UCD.

'Finding proper ways to connect the real and virtual geographic worlds that the user of geo-mobile applications interfaces with, could be one of the keys to developing more usable geo-mobile applications. One result that could directly contribute to making guidelines for more usable geo-mobile applications is the determination of the types of LMs that support user geo-identification and that should be included in mobile map interfaces' (Delikostidis and van Elzakker, 2009). LMs are distinctive stationed objects of real worlds which can be used in finding relationships between locations of objects and the paths for navigation purposes (Sorrows and Hirtle, 1999). There are several categorizations regarding the aspects which make a LM distinctive in literature such as attractivity, memorability, semantics, functionality, etc. Many current navigation systems just utilize turn-by-turn and distance instructions as navigational aid, which is not the best navigation solution due to the natural human ability in wayfinding via LMs (Brenner and Elias, 2003). Likewise, some current systems present LMs non-selectively, which is also not proper e.g. due to imposing high cognitive load on the user or the low rendering process (Delikostidis and van Elzakker, 2009). Currently, there is a debate in literature about the use of proper selective LMs in (pedestrian) navigation systems. Considering these arguments, selection of LMs is used for evaluating different methods in this case study.

A usability testing methodology for mobile geo-application can be divided into four stages (Delikostidis, 2007). They briefly include defining the following items:

1. (a) Research questions and objectives, (b) Using field-based or laboratory-based testing.
2. Prototype being evaluated (real one or simulation) based on previous stage.
3. Variables that are going to be tested and how to measure them.
4. (a) Techniques of user data collection which leads to measuring the variables, (b) Defining environmental conditions.

The above items for the present case study are realized as follows:
1. (a) The main purpose of this case study is to compare methods and techniques for the usability evaluation of a geo-application in the requirement analysis phase of UCD. The main subjects in evaluating different methods here are the LMs. According to these objectives the following questions would be relevant to be answered.

- Which of these methods is the most efficient, effective and satisfactory for this context of use?
- Should ET be used alone or along with other methods (mainly TA)?
- Could the qualitative and quantitative data (what users mention as selected LMs and their ET data) confirm each other? This would be an extra effort on investigating the question: 'Does ET and more conventional usability technique like TA, in geo-domain always support each other?'
- Does TA affect eye movements and how?
- Questions related to LMs are:
    - By selecting LMs, what information is obtained by ET that cannot be obtained by other methods?
    - How many of the presented LMs on the PDA were noticed in the environment?
    - How can I verify that objects looked at by participants are considered/used as LMs?

(b) A field-based usability testing is selected for this case study. There are two reasons for this selection: 1. Due to recreation of the real context of use and users interacting with the real environment, field-based testing is more preferable than laboratory-based usability testing, although it requires more resources. 2. Mobile eye trackers are mainly designed for outdoor application, a better assessment of their usability can be achieved in their actual context of use.

2. As mentioned before, the prototype being evaluated is a PDA, i-mate ultimate 9502 running iGO 2006 software which is a commercial product.

3. Standard variables of usability testing are efficiency, effectiveness, and satisfaction. For this case study, the efficiency of the 2 different methods can be determined by measuring the required time for data collection in each method. Effectiveness can be assessed by measuring the extent of received information by each method. Satisfaction (of the researcher) which is a more qualitative measure compared to previous ones, can be measured through the assessment of the friendliness and complexity (or simplicity) rate of applying different methods. This includes the ease (or difficulty) of applying different methods before the test in the preparation stage, e.g. in doing calibration. It can also be assessed by rating the complexity (or simplicity) of use during the test, e.g. the number of data losses due to calibration or technical problems. Finally, it can be measured after the test by assessing the ease of data analysis of different methods.

4. (a) The applied techniques for user data collection in this case study include ET, TA, questionnaire, and drawing mental maps. The applied techniques are described in section 3.2.1.4 in detail. (b) In a properly designed (field based) usability testing, the environmental condition should be the same as the real world's context of use. An actual user may use the PDA for a navigation task in various environmental conditions including different times (of the day), locations, or weather conditions. However due to some limitations in applying data collection devices (e.g. sensitivity to water), the

surveys have to be scheduled based on some environmental condition. These conditions are described in section 3.2.1.2 in details.

The next sections define other specifications of this case study including test participants, test environment and the study area, test procedure and the applied method and techniques for implementing the scenario.

3.2.3.1. Test participants

The test participants are usually selected among the actual target groups of users. Using a skilled user or using an underqualified test participant will bias the outcomes of a usability testing. For this case study, the participants were chosen among ITC students. According to (Streefkerk, 2006) students can be used as test participants provided that they are part of the actual user population. We will try to choose the participants among students with different levels of experience with GPS and mobile maps, such that they can evenly be distributed among two groups of participants. Later, a method will be assigned to each of the groups and then, the user data resulting from these two different methods will be compared to each other. ITC students mostly have a background of using maps or other related geo-applications and this may bias the results. Still they are part of actual end users of a mobile geo-application and they are accessible for implementing the test. Moreover, the objective of this case study is to compare different methodologies related to each group and participants are assigned evenly among different groups. This implies the amount of bias for the 2 groups is rather the same value. So, it can be ignored.

In the UCD approach the number of participants differs according to the type of technique being used. Usually in the requirement analysis phase, the type of usability testing is more qualitative compared to the final usability evaluations which produce statistical results. As a consequence, the number of required participants for the usability testing at the primary stages is typically less than required people at the final stages of the process. For a usual qualitative technique a minimum of 4 to 9 test persons are required. According to (Pernice and Nielsen, 2009) for an ET usability testing (which does not require to produce a heatmap) using 6 participant is appropriate; even the acceptable data from 3 users are sufficient. Since we have 2 groups of users with different combinations of usability testing methods applied to each group, using 6 users in each group, which leads to a total number of 12 users would be a reasonable choice.

For selecting participants such that they can form 2 homogeneous groups, a pre-selection questionnaire for demographic information of the participants will be distributed among ITC students. A sample of this questionnaire is included in the Appendix B. The questionnaire asks for general personal information of participants, their familiarity rate to the study area and their amount of knowledge and experience with geo-applications including paper/digital maps, GPS and navigational systems. Each participant will be referred via a code like A1, B1, etc. Any information regarding users' backgrounds and their actions during the test is kept confidential, and they may only exposed using these codes.

3.2.3.2. Test environment and the study area

The test environment is an important factor in designing a test. Usually the test environment in a usability testing is chosen in such a way that it is similar to the actual environment in which end users are going to use the product. Type of end users, test purpose, product's type, test equipment and the stage of testing the product are the decisive factors of selecting an environment. For the purpose of this case study, since the users are required to navigate towards a predefined destination while selecting LMs we tried to define the environment such that it includes a variety of LMs.

Different levels of familiarity affect the users' selection of LMs (Najari, 2009, Xia et al., 2008), hence for selecting salient LMs it is preferable that users are unfamiliar to the study area. As mentioned earlier the main focus of this case study is to assess the usability of the ET technique by comparing 2 different methods and for this purpose it is tried to form 2 user groups with homogeneous demographic and expertise backgrounds. Another factor which is regarded in forming homogeneous groups is the familiarity level with the study area. Since the aim is to compare the methods of the 2 groups and both groups have rather the same familiarity rate with the study area (familiarity rate to the study area was one of the questions of pre-selection questionnaire), the bias is ignored. Considering the resources and the time available for this research, the determined study area for this case study is part of the center of Enschede which can satisfy the above condition (figure 3.5).

There are also some other considerations regarding the test environment due to the equipments' specifications used for user data collection. For instance, the eye tracker and the PDA are sensitive to water and do not work if they get wet. Also, the eye tracker scene camera does not provide good contrast in dim lighting conditions. So, the experiment should be scheduled in brighter hours of the day or when it does not rain or snow. Moreover, due to seasonal weather conditions, if it gets very cold the participants may refuse to attend the test.

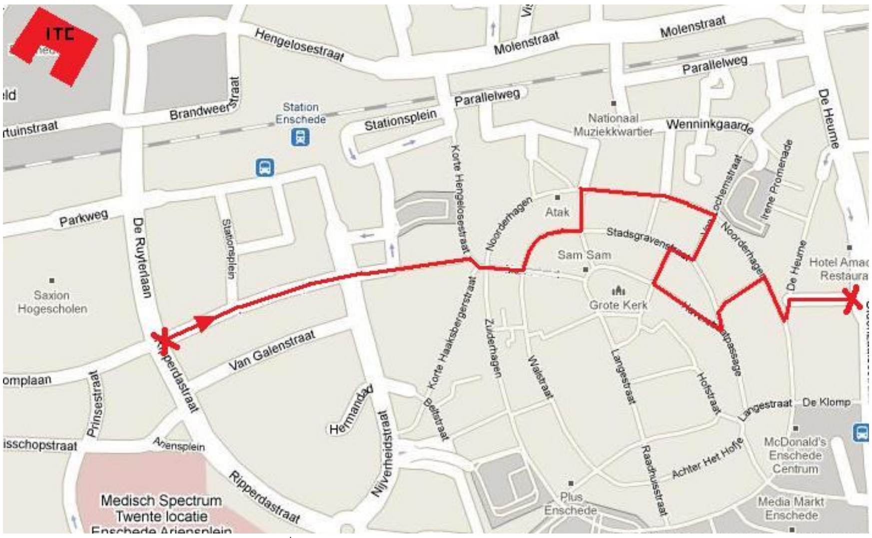


Figure 3.5: The study area for the 1st case study with highlighted predefined route (about 1 km).

3.2.3.3. Test scenario

In order to achieve research objectives, a pre-defined task based scenario is prepared in which two groups of users are participating. For each group, different user data collection methods are used which include using the ET technique alone and ET along with TA. During the test and while the user is navigating, the researcher walks all the time at a reasonable distance behind the participant and in the case that user misses the correct path and makes mistake in following the predefined route, or there is a problem, the researcher approaches the user to guide him.

Before the test, users in each group fill in a questionnaire in order to collect their demographic data and their back ground information regarding maps and navigation. Likewise after the test, participants take part in a semi-questionnaire in which their opinions regarding the experience are collected. They are also asked to draw a mental map of the path. To do the experiment for each group 6 test persons are required which makes a total number of 12 participants. A rather short and predefined path, which should also include different types of LMs, from a known starting point like 'A' to a known destination like 'B' on the screen of a PDA is identified (highlighted). These 2 points are pinned on the map using one of software's functions. The applied map layers in the PDA for this test for both groups are identical, including roads, their names and all the main LMs along the path. Participants are required to wear the eye tracker and to navigate exactly on the predefined path from 'A' to 'B' using the PDA. Also, depending on the group number, the users are required to do TA or just navigate without TA (table 3.1). The following is a stepwise description of the test scenario:

Table 3.1: Summary of applied methods and groups.

	GROUP 1	GROUP 2
METHOD ADOPTED	Questionnaire	Questionnaire
	Thinking aloud & Eye tracking	Eye tracking
	Mental map & Semi-questionnaire	Mental map & Semi-questionnaire

As mentioned earlier, before starting the test, 2 homogeneous user groups are selected through a pre-selection questionnaire. According to the determined time schedule for each participant, on the day of the test the researcher and the participant meet at the ITC usability lab. After a brief introduction to case study, the task and the study area, the PDA is given to the user and he is given some time to familiarize himself with it or to ask questions. Also, the only allowed functions of the navigation software which are required for this task (zooming, panning and rotation) will be explained to the user.

Then a calibration for 3 meters distance and using 5 to 9 points is done in the ITC lab, and the related instructions in using the eye tracker are provided to the user (e.g. users were required not to touch the glasses at all during the test, and not to make a sudden or strong movements). As the next step, the researcher and the participant walk to the pre-defined starting point, however, the user is asked to practice immediately after leaving ITC and before reaching the starting point as a preparation for real test. At the starting point, the researcher records the time and the participant is then asked to orient himself and start moving based on the predefined path and (depending on his group) do or do not do think aloud as he walks towards the destination point. In the case of using TA the participant is asked from the very beginning to think aloud and mention whatever comes to his mind regarding the navigation process as much as possible. The statements could be about whatever he reviews in his mind for finding his way, the attractive stationary elements or what else he finds to be proper as LM in the environment, no matter whether it is a familiar item or a new one to him. Due to the sensitive nature of eye movements, the first group will be asked to verbalize the 'name' of the selected LMs with just a word or phrase like 'that building', 'this junction' or just 'this one' (the intention of TA in this case study is just to confirm and differentiate between what the user mentions and what he is looking at, but does not mention, with the lowest possible disturbance due to TA to the eye movement patterns).

As the participant is completing the task, the researcher follows him at a reasonable distance and takes notes of his different reactions. At the destination point, the researcher records the time and the participant can remove the devices. After that a semi-questionnaire is carried out to acquire extra information about user's impressions and experiences with the ET equipment and TA technique (Appendix C). Also, drawing a mental map of the path which the user navigated through including LMs, confirms the main recalled LMs by the user after the test.

Some clarification regarding the method

- Since the ET method for selecting salient LMs has not been experienced before, it is not easy to predict what the outcomes would be. Nevertheless, it seems 'when a user is talking, he may look for a longer time on an object'. This can be states and tested as hypothesis like 'fixation durations in a combined ET and TA method are longer than applying ET alone'. Also, it is expected that these 2 different methods (ET, and ET with TA) provide different user information for selecting salient LMs for the design of a pedestrian navigation prototype.

- The assumption for the analysis of ET data is, that when a user selects a LM, he spends more time looking at that LM than he spends on another object present in the environment. If this hypothesis is true, we may say, for instance, that there is a time threshold, and when the time that users spend to look at an object exceeds the threshold, the object most probably can be regarded as a LM, even if user did not mention it in TA. This idea can be used to find what extra information (here: LMs) ET can add to the already applied methods (e.g. audio-video recording) for selecting LMs.

- Participants in group 1 who apply TA technique are asked to use a ‘short phrase or word’ for mentioning the LMs. Eye movements have a sensitive nature and any distributing factor (e.g. additional visual information or audible instruction) can interfere the state of the point-of-gaze (Hayhoe and Ballard, 2005). Using short phrases for mentioning the LMs is to minimize the disturbing effect of TA on ET data. It makes it possible to compare ET data of the verbalized LMs with those LMs which are extracted by calculating the duration of the spent time on watching them, although they are not expressed as LMs by user. The researcher should explain to users that because of the presence of an eye cursor we already know what they are talking about, so there is no need to add extra explanation about the LM.

- In mobile ET, in contrast to stationed system, the background changes constantly. Also, due to free head and body movements, the participants’ behaviour and eye movements are very inter-individually variable. In order to apply mobile ET, consequently, it is necessary to use short tasks where inter-individual similar eye movements can be expected (Mayr et al., 2009).

- The reason for highlighting the path is to prevent that users get lost. Since the route is predefined and also it is not a straight line, it would not be easy for user to remember it.

- According to (Pernice and Nielsen, 2009) for a typical (stationed) ET usability study, around 6 participants would suffice. For finding common LMs among users and comparing them, having repeated similar eye movement data is necessary. It is expected that the data in mobile ET are inter-individually variable. Hence, for having repeated similar eye movement in mobile ET maybe more participants are required. Furthermore, this is the first experience with the system and potential problems regarding data collection (e.g. the system not recording data, poor calibration, etc) are predictable.

3.2.3.4. Test techniques

The techniques which will be used in this research include ET, TA, (semi-) questionnaire, and mental map drawing. ET is the main technique which is going to be used in this case study. TA, questionnaire and mental map are the accompanying methods which depending on the users’ group will be added to the main technique. In summary:

- ET is the main technique which is going to be investigated in this research.

- TA is an important technique in giving insight into the user’s thoughts by expressing their actions during task execution. Although it is a rich source of information which cannot be replaced by other techniques, the intention here is to assess its negative and positive effects on ET by comparing 2 methods: ET alone and ET with TA.

- A questionnaire is applied before the test to collect users’ demographic data, and to organize 3 homogeneous groups of users regarding their background and experiences for comparison purpose.

- A semi-questionnaire, or a combination of questionnaire and interview, is applied at the end of the test to complete the provided information during the test and to obtain extra user information regarding the test. The intention is to gain more insight into user’s recalled information after the test. During the interview, those users who may have limited writing or drawing ability will talk about the experience while it is being recorded by researcher. The semi-questionnaire will be applied to both groups.

- Mental map refers to what a participant perceives and keeps in mind regarding some locations to which he attended before. It can be achieved by asking the participant to draw a sketch of the route (and important elements he remembers) which he navigated before reaching the destination. It is a proper technique which confirms what the participant really perceived and remembered as LMs.

3.3. A case study with a fixed eye tracker

3.3.1. Introduction

The previous section (3.2) described the design process of a case study with a mobile eye tracker. In this section, we describe the same process but for a case study with a stationary system. Like the previous case study, first we explain the objectives of designing the test, the prototype to be evaluated, the applied eye tracker for the test and the encountered problem using the equipment. Then other specifications of the test, like test participants, environment, scenario and techniques are described as well.

3.3.2. Stationed-ET user research methodology in the evaluation the design evaluation phase of a geo-web application

The aim of this case study is to obtain information about the application of the (fixed) ET technique in the geo-domain. In this case study a geo-web application (which is a prototype for visualizing iceberg data) is evaluated, using ET and some other usability techniques to check for the potential problems of using the prototype. The prototype is developed by Nguyen (2010) in the context of MSc research (see section 3.3.2.1). Considering that it was designed in a short period of MSc research, and the fact that this is the first usability evaluation carried out on the prototype, it is expected that there are potential problems which can be exhibited by usability testing. In order to realize the main objective of this case study regarding the usability of the fixed ET technique in providing extra information, like in the first case study, two main methods (ET alone and ET combined with TA) are used to compare the resulting user data. This case study is a joint attempt (of the researcher and another MSc student, Nguyen) in using ET from 2 different points of view. The objective of the study, not relevant to this research, was to assess the usability of the prototype itself and its design solution for visual exploration of iceberg data. The next section describes the characteristics of the prototype which we used for the test.

3.3.2.1. A prototype for the visualization of iceberg data

As mentioned, the prototype is developed by Nguyen (2010) with the aim of enabling visual exploration of trajectory data on the web, using existing web visualization libraries (figure 3.6). To implement this prototype, first the iceberg data are stored in a PostgreSQL data base, then PHP was chosen to build the middle ware between database and client side and finally, the GUI of the prototype was created by using ‘Timemap.js’. The prototype is mainly designed to provide answers to the questions of ‘What’, ‘Where’ and ’When’ some events (e.g. (dis)appearance, split, movement) happened to icebergs. In the prototype there are 2 data file used: the first one only contains the starting points (appearances) and end points (disappearances) of icebergs, with straight connecting lines between start and end to indicate the major directions. The second file contains all observed iceberg positions, also at intermediate locations, and the connecting lines to more accurately represent the trajectories. All the events belonging to the same iceberg are displayed with connected lines. At low zoom level (overview) the first data file is displayed, at high zoom levels (details) the second data file is displayed. The prototype contains 4 main sections which are also considered as 4 ROIs (see section 2.6.3) for measuring gaze data: legend (left), functions (right), timeline (middle-up) and map (middle-down). The map view and the timeline contain information of all events and they are linked together (e.g. by clicking on an event on the timeline, the map view will pan to the position of that event and also the pop-up window of the corresponding event on the map view will be shown). The prototype only displays those events on the map view which are in the visible range of the timeline. There are two bands on the timeline: the top band’s time interval is set in months. The bottom band's time interval is set in decades to give the users an overview of all the data. Timelines can be dragged forwards or backwards using the mouse. The prototype also provides (filtering and animation) functions. The icebergs can be filtered by name, size, lifetime, average speed and travel distance. The jump to time function lets users go to any time on the timeline without dragging it or using animation functions. Finally, the prototype provides some animations functions. Due to the temporal irregularity of iceberg data, 2 sliders are created to control the speed of the animation. The first slider controls the speed of the change between 2 scenes. The second slider controls the time interval step between 2 scenes (e.g. when users set the time interval step to 30, the next scene will be the scene of 30 days after the previous one). To reduce the speed problem in running the prototype during the test, the iceberg data base (from 1976 onwards) was reduced to a data set from 1979 to 1988.

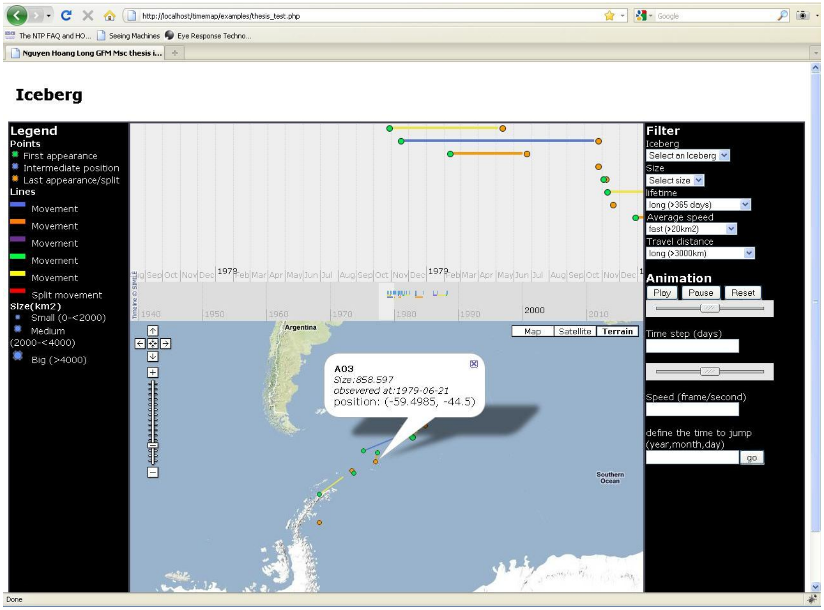


Figure 3.6: The prototype (for Antarctic iceberg data visualization) to be evaluated via the ET technique (Nguyen, 2010).

3.3.2.2. Test equipment: faceLAB Eye tracker

In this section we first explain characteristics of faceLAB Eye tracker and then we describe the encountered issues in using the system.

System specification

faceLAB ™ *4.5* is a stationed (VOG based, 60 HZ frequency) eye tracker by *Seeing Machines*, which is a technology company with a focus on vision based human machine interfaces. The Seeing Machines technology is based on computer vision processing technologies that allow machines to track human eyes and facial features. The faceLAB software uses a set of cameras as a passive measuring device. Images from cameras are analysed to work out characteristics of a subject's face, including the current position and orientation in 3D space, the gaze direction and several other measurements. In the calibration process, the system first makes a head model of the subject. The head model is the way of remembering how features of a subject's face look and relate to each other. It can be saved and re-used later. Head tracking is the process of finding feature templates and determining the head-pose using the position of these templates, and it is shown on the screen by an overlay onto the video display (figure 3.7). The main output for data processing from head tracking is the 3D head-pose. It can be used to differentiate head movements from eye movements, which allows a more robust gaze tracking and leads to more freedom of the user.

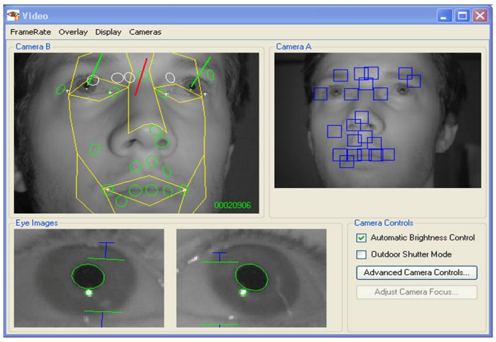


Figure 3.7: Illustration of created feature templates (head model) by faceLAB, overlaid on the video display.

The system originally is composed of the following hardware and software platform: HP laptop (or PC) for the calibration and data recording (by faceLAB software), stimuli display and analysis (by gaze tracker software, version 07.02.251.192 full for faceLAB, which calculates the fixation data, received from the faceLAB, and adjusts it with the stimulus on the screen), a pair of point Gray Flea (or SONY) cameras, a Stereo-Head which holds the 2 cameras, designed to be adjustable for different user configurations and 2 infrared pods for precise gaze tracking and for tracking at night or in a dark environment; with the flea cameras a third IR pod is also provided (figure 3.8). The IR pods are powered off the stereo-Head cable or by a separate power cable. Setting the IR pod positions is done within the IR pods position wizard and different camera configurations require different IR pod configurations. (E.g. good placement of IR pods for *precision-gaze tracking* produces a triangle of reflections in the subject's eyes – like what ASL mobile eye does). The 3 IR pods emit the IR light towards the subject's eyes and face and the reflection is captured via 2 cameras which determine the head model as well as the location of the iris and pupil. The calculation of gaze utilizes the head pose and the location of the iris and/or pupil (whichever can be detected more reliably). The system is not intrusive in some senses, for instance when the subject moves his head, software can recommence tracking whenever he returns and his head is visible to both cameras. However, for raising the accuracy it is recommended that user prevent extra movements. The typical accuracy of gaze direction measurement is 0.5-1° rotational error, and for head measurement within +/- 1mm of translational error and +/- 1° of rotational error (the error of calculating gaze direction without calibration is up to +/- 5°).

Figure 3.8: The faceLAB eye tracker with a pair of Flea cameras and three IR pods.

faceLAB is compatible with the gaze tracker analysis software. Although it is possible to run the whole set-up on one system, in the system we used, the set-up was split on to 2 computers (figure 3.9): a laptop for doing the calibration and recording data (running faceLAB), and a PC for the stimuli display, the analysis and saving the data (running gaze tracker). The reason for a 2 computer set-up was that in the case of any crashes to the laptop, the data on PC will remain safe. The idea is that the subjects use the PC for running the stimuli and performing the tasks, while the experimenter can check the calibration on the laptop during the whole test (The experimenter cannot see user's screen on the laptop during the test). There was also a (keyboard/screen/mouse) splitter for switching the views of the user after calibration.

Figure 3.9: The applied faceLAB eye tracker running the stimulus. The user's monitor is captured via an external camera as back up.

Calibration issue

One of the encountered problems in using faceLAB was the calibration issue. The calibration did not provide the same results every time. On average, we did the calibration 3 times for each person. We checked the accuracy of calibration before starting the test for each person via the software. The software provides the possibility of seeing the eye-cursor in real time for checking the accuracy by asking the user to look at some particular points on the screen or to follow the mouse. We noticed that the accuracy differs on the different parts of the screen, and it was not the same amount for each user. Also, the amount of noise (jumping of eye-cursor) when the eye was fixed was more than when the eye was moving (e.g. following the mouse-cursor). In data analysis, the lack of sufficient calibration accuracy (e.g. to distinguish which functions the user is looking at, at each moments) is a factor that limits detailed analysis of user data.

Another problem is the effect of calibration accuracy on defining ROIs (lookzones). E.g. the accuracy was not enough to distinguish 2 close functions from each other; hence big ROIs were defined which makes detailed analysis impossible. After finishing the test some obvious systematic shifts were removed via software, but for the random shifts, we did find a solution. In order to do this, we considered some of the obvious objects of the prototype which a particular user used and fixated on them frequently, e.g. the zooming function of the map zone, or the functions of the function zone. Looking at such objects, we noticed that for some users there was a systematic shift for the location of fixations on the prototype. Then we tried to estimate the amount of correction visually (e.g. 0.5 cm vertical and 1 cm horizontal shift), and adjusted it with trial and error via the software (e.g. figure 4.10 group1, user B1).

Hardware issue

The PC (Pentium 4, CPU 3 GHz, RAM 1.5 Gb) on which the gaze tracker run was not powerful enough to run the gaze tracker software and the application at the same time. Gaze tracker is considered a heavy software and it needs a lot of computer's resources. While gaze tracker was running, the application could not run smoothly. This resulted in slowness of the system when a participant was using the application. Since eye movements are so sensitive, the slowness of the system will certainly affect the eye movements of a subject. E.g when the subject clicks on a function, he will look at other parts of the screen while he is waiting for the system's response. This makes it hard to compare 2 subject's eye movements. Another big issue was that the system's slowness was random, so ET patterns were not recorded in an even condition for all users. It means for one subject the system was really slow, while for the other subject it was not that slow. When the system was slow, the user had more time to look around the webpage. This problem makes it hard to assess the factor of *efficiency* (e.g. sometimes the user was able to easily give the answer, but he had to wait because the system was slow, and sometimes not).

Because of hardware problems, data analysis and exporting the resulting avi files (videos) took a lot of time; the system could not respond many times and we had to restart.

Software issue

The resulting video-data is not complete and this makes the analyses too complicated. There are 2 different output videos for a test. One is a simple screen logging of the prototype including audio recording, mouse movements and changes of the prototype during the test. The other video output contains eye-cursor over the stimulus, however due to hardware problems it did not record the changes of the stimulus properly, so we cannot see what is really happening (e.g. when the user clicks on a combo box, sometimes the recording shows that a combo box is opened and sometimes the combo box does not change at all). It also did not record voice or the mouse cursor and the resolution is low (we lowered the resolution to raise the saving speed). The only solution to analyze these 2 videos is to synchronize via a software. Even after that we would have to look at 2 cursors (eye- cursor and mouse-cursor) at the same time to see 'what is going on' and to remove doubts.

Other issues

During test execution we noticed some other general issues which removing them improves the resulting data.

- To deliver task-related questions during the test to the user, we had 2 options: to give them the questions verbally or in a written way. We could read the questions 1 by 1, which was not a good solution because during the test, the user sometimes looked at a question several times to understand it better, so they needed to see the questions. To give the questions as text, we first decided to simply display them in a text file on the desktop, so the user can open it as an extra window and switch between windows (application and text file). Due to the system's slowness and the fact that it may affect eye movements, it was not also a good idea. As a solution we wrote down each group of questions (see 3.3.2.5) on a piece of paper and fixed that near the monitor, however the user had to move his head to look at them; this may reduce calibration accuracy. A better solution may be to use a bigger size computer screen, such that the stimulus and the text files can be displayed at the same time. This solution was suggested by 2 users during the interview after the test.

- The participant's chair had wheels and it was not stationary. During the interview some users complained about that and said that the chair made it difficult for them to stay in a stable position during the test.

Finally, as mentioned in the 1st case study choosing an eye tracker depends on our available facility, and the specific application it is used for. According to literature, faceLAB is one of the commonly applied eye trackers among other stationed systems. The major reason for analysis failure in our case study was due to a hardware problem which can be solved. However, comparing faceLAB more accurately to other famous fixed systems like Tobii in practice requires more investigation which is not possible during this research.

3.3.2.3. Test participants

In this case study there are 2 groups (ET alone and ET combined with TA) of 4 participants, which make the total number of 8 users. As mentioned in the previous case study, in usability testing, participants should be selected among the real target groups of users. The actual end users of such a prototype (visual exploration of icebergs) usually are the experts and iceberg specialists who have some knowledge and background about the icebergs. However, since our objective is to compare different methodologies to collect user data (like in the previous case study), and participants are assigned rather evenly between the 2 groups, we ignore the bias. Hence, participants are selected among ITC students and staff (with more or less similar background information about the icebergs and trajectory data) via a pre-selection questionnaire (Appendix E). Also, we noticed in the pilot study that the system is sensitive to different eye conditions. E.g. the accuracy for a user who wears contact lenses was not acceptable, although for another person who also wears contact lenses it was satisfactory. We also noticed that the system worked with a person who wears glasses with acceptable accuracy, while for another person who wears bi-focal glasses, it did not work at all. These results were considered in selecting participants via the pre-selection questionnaire.

Usually for an ET usability testing around 6 persons are required, however according to (Pernice and Nielsen, 2009) even 3 acceptable ET data for a qualitative experiment suffice. For this test, there are 2 groups of participants and each group includes 4 persons who used the same method. This makes the total number of 8 participants.

3.3.2.4. Test environment

The test environment in a usability testing should be similar to the real environment of use. For the purpose of this case study the environment was an ET lab at the UT (Twente University). It was a small room (with normal lighting condition) including one laptop and one PC (eye tracker) linked to 2 cameras and 3 IRpods. Since the user of such a prototype (visual exploration of icebergs) usually works in an office or lab similar to this room, the environment for testing the prototype was appropriate, except that the user was restricted to not to move too much.

3.3.2.5. Test scenario

On the day of the test, according to the time schedule for each participant, the researcher meets the participant at the ‘Hallenweg’ busstop in front of the ET lab building at the UT. Then a brief introduction to the aim of the case study, the ET (and TA) technique, the calibration procedure and the prototype under evaluation are provided. In order to provide the same levels of information to all users, a written description of icebergs characteristics and the functionalities of the prototype is then provided to the user, and he is allowed to ‘play’ with the software for around 10 minutes and to familiarize himself with it; he can also ask questions about the whole procedure. Then the calibration is carried out. To do the calibration, after running the faceLAB software, the first step is to ask the user to adjust his position using the face image provided (by faceLAB) on monitor screen in front of him. The researcher helps the adjustment by measuring the user’s head distance from the screen, and checking the acceptable distance range with the software. Then the user is asked to look straight and to remain still for a few second, while his head model is created by the software by clicking the set model button. The next step is to calibrate and adjust the gaze relative to the screen. To do this, 9 lighting up points appear 1 by 1 on the screen and the user is required to fixate on the center of each point while they are lighting up; after this the eye tracker knows where the user is looking. Finally, we start data log-in, first in faceLAB and then in gaze tracker on the other computer (after running the application on that system).

There are 9 predefined questions which require the user to apply almost all views and functions of the prototype. The questions are prepared by the prototype developer and are organized into 3 groups of 3 questions related to main user tasks, namely overview, filtering and zooming, and details on demands (Appendix F). In the case of using TA, the user is required to mention what he is thinking while carrying out a task. Although the software logs together with the stimulus, a video of the monitor is captured during the whole test as a back up via an external camera (figure 3.9). Each group of questions is presented to the subjects on a piece of paper which is fixed on the left side near the monitor (after finishing the first group of questions, the paper is replaced by researcher with the second group). In order to know when the task execution starts, the user is asked to first read each question aloud. Then when he finds the answer for each question, he will say aloud the answer. At the end of the test the gaze tracker stops data log-in by pressing F2. Finally, a semi-questionnaire for obtaining extra information regarding the objectives of the research is provided to the user (Appendix G). The summary of applied techniques for each group is presented in table 3.2.

Table 3.2: Summary of applied methods and groups

	GROUP 1	GROUP 2
METHOD ADOPTED	Questionnaire	Questionnaire
	Thinking aloud & Eye tracking	Eye tracking
	semi-questionnaire	semi-questionnaire

3.3.2.6. Test techniques

The applied techniques in this research include ET, TA, audio-video recording (via an external camera or screen log-in) and (semi-) questionnaire. ET is the main applied technique in this case study. TA and (semi-) questionnaire are the accompanying methods which, depending on the users' group, will be added to the main technique.

- ET is the main technique which is going to be investigated in this research.

- faceLAB software provides screen log-in with the prototype (including mouse movements and prototype changes) along with synchronized audio recording. The same information can be obtained via an external camera. Since the eye trackers sometimes fail to record the data, and due to the fact that this is our first experiment with the eye tracker, we decided to use an external camera as a back-up.

- TA is a powerful technique to get a deeper insight to what a user thinks and what problems he encounters while completing a task. Like in the previous case study, the intention of using TA here is to assess its negative and positive effects on ET by comparing different methods.

- A questionnaire is applied before the test to collect users' demographic data, and to organize 2 homogeneous groups of users regarding their background and experiences for comparison purpose.

- A semi-questionnaire is applied at the end of the test to complete the provided information during the test and to obtain users' opinions, complaints or satisfactions in applying different techniques (ET and/or TA).

3.4. Conclusion

In this chapter we described the design stages of the 2 case studies with 2 different types of eye trackers (fixed and mobile) in order to investigate and experience the usability of the ET technique in the geo-domain in practice. First, for each case study the aims of designing the test, the applied equipments for the test and the encountered problems in using the equipments was explained, and then other specifications for designing the tests, like environment, participants, scenario and techniques were described. The next chapter will describe the test execution, the outcomes and recommendations for the improvement of the method for each case study separately.

4. Test execution and the outcomes

4.1. Overview

In this chapter, we explain the actual procedure of executing each case study, based on earlier mentioned methodologies. Then the analysis of the data, the encountered problems during test execution and analysis of the data and finally the obtained results of this research will be described.

4.2. The first case study: mobile ET

4.2.1. Execution of the test

The test was carried out from 15th to 19th of February 2010. Before executing the real test, a pilot test was held on 12th of February with one of the ITC students as participant, during which ET and TA was tested, to check the quality of video, sound and the whole test procedure, in the real test environment. During pilot testing it was detected that the sound quality of Mobile Eye was not clear enough in a real environment to be analyzed. Due to this, an external microphone was applied for the first group in the real test. Every day, the tests started at 9 AM and finished at 5 PM. Each participant did the test individually. The average spent time by a test participant for the whole test was about 1:15 hours. Every day 3 users participated in the test; because of weather conditions and the temperature it was not possible to handle more that 3 user per day, and after finishing each test at the end point, the researcher had to walk back to ITC to meet the next user. The analogue ET data of each user was recorded on tape. Since data transfer takes the same amount of time as the test, the data for all users was saved digitally at the end of each day. As mentioned earlier, users were selected based on a pre-selection questionnaire in a way that they form 2 homogeneous user groups and each group uses a different method (figure 4.1).

(a) (b)

Figure 4.1: (a) A participant of group 1 who uses the Mobile Eye and an external microphone, (b) A participant of group 2 who uses just the Mobile Eye.

The test originally was executed by 14 users (plus 1 pilot), but some of the data was lost due to the following problems:

- The main issue which made data unusable was the 'calibration loss' and there were a few reasons for that. After finishing a test and checking the data it appeared that for some users although the calibration was accurate enough at the beginning of the test, it gradually reduced and data became totally unusable. The Mobile Eye spectacles are heavier than usual glasses, and it is normal that the glasses move a little during the test, while the user is walking. There is a headband attached to the glasses which can be adjusted and tightened for different head sizes (figure4.2). The design of adjustable headband is not good enough to secure the glasses properly, especially if the person has a small nose it moves more. After detecting this issue we used extra clips to tighten the glasses for the remaining users. This reduced the amount of shift significantly.

Figure 4.2: The adjustable headband of Mobile Eye.

- Furthermore, because of the low temperatures, most users needed to wear a hat and scarf. We did our best not to touch the glasses and the attached wire (SMU) while users were wearing the hat and scarf after the calibration and during the test, but still this issue caused some data calibration problems.

- One user's data was lost after the test, due to an error of the eye vision software while saving the data. The calibration information for each user is saved in an evi file, which should be loaded before saving data. Because of an error of the software, the evi file was corrupted during saving.

The remaining acceptable data are for 8 users (4 users, each method), which still creates 2 rather homogeneous user groups. All participants are from GFM MSc students. The average age of participants was 29 and they were from different nationalities. The summary of background information for the 8 subjects is compiled in table 4.1.

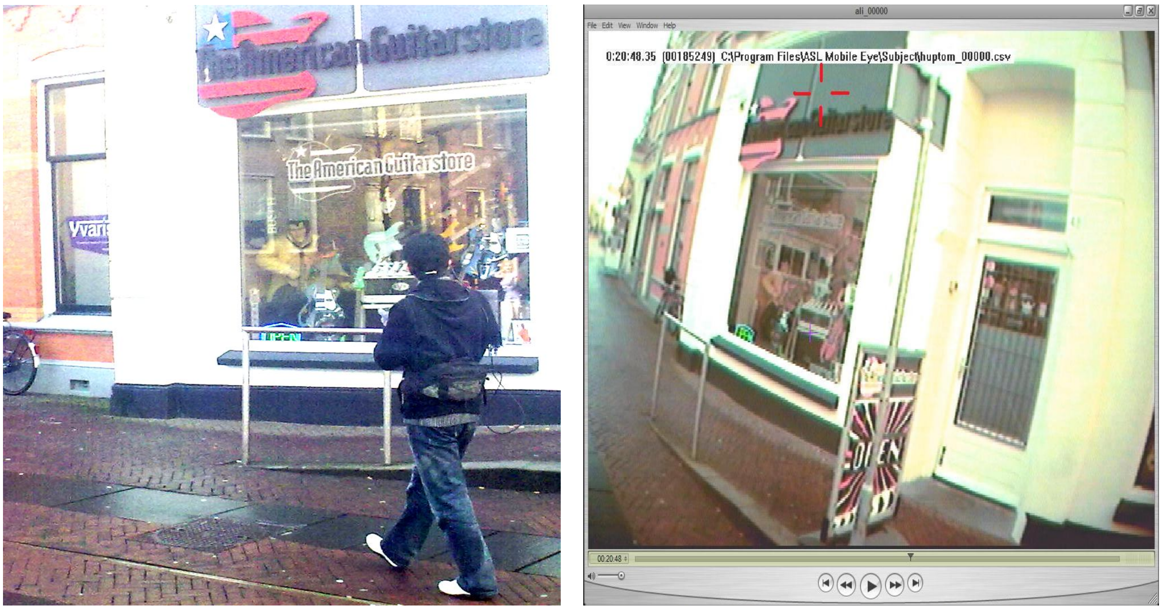
Table 4.1: Summary of background information for each user group.

Familiarity rate with: (percentage)	Group 1 ET-TA	Group 2 ET
Paper map	90	90
Digital map	65	90
GPS	65	65
PDA	65	40
Navigation systems	65	40
Study area	90	80

The other problems were about the temperature, which (despite using gloves) forced some of the users to stop the test for a few seconds to warm their hands for handling the PDA. There were other cases in which people (e.g. shop or restaurant owners) asked us to move and do not record data there. In another case, the police interrupted the test by asking the user about the test. In each case, the researcher explained the case to them, so that the user could continue the test.

4.2.2. Analyzing procedure and results

The resulting avi data files of the Mobile Eye include the scene (environment) and the point of gaze (eye cursor) overlaid on the scene (figure 4.3).



(a) (b)

Figure 4.3: (a) One of participants is looking at a landmark during the test, (b) The recorded scene with the overlaid eye cursor (red cross) for the same participant at the same time.

The created files were large (e.g. 10 Gb), and the host PC was not powerful enough (Pentium 4, RAM 1 Gb, Hard disk 80 Gb) to display them smoothly. In order to transfer these files from host PC to other systems with the same quality, we had to split the files with some software (e.g. HJ-Split) and to merge the files in another system. Analysis was manual, since no access to ET software. In order to analyze the data, we had to watch the videos in a slow motioned mode to extract the time that users looked at LM's. Some of the usual video players were checked and *Apple quick time player* was chosen to analyze the videos. Because of the file size, other players could not display the slow motioned videos properly. In order to analyze the data, first all the main LMs on the pre-defined path were identified and coded (Appendix D). This included almost all real worlds' stationary objects on the path with at least one discernible attribute related to their functionality, shape, size, color, etc. Then for each user, the spent time on each LM was extracted and written down under the related code. The LMs which the user named by TA were recorded separately under their codes. If after looking at a LM, the user looked around or turned and then looked at that LM again (i.e. the user looked at the same LM several times), the durations were added. But if the user missed the route and he decided to go back to a previous point and start from there again, the values for the repetitive path were not included in the calculations. The calculations were a very tedious and time consuming procedure, since the path included around 100 main LMs. The average of total number of the noticed LMs by each user for group1 was around 95, and for the group 2 was about 75. Some sections of the videos were watched twice or three times in order to obtain the correct durations. The eye tracker provides the time information (in the left upper corner of the video frame) with an accuracy of 0.01 second. However, due to time constraints and lack of required software, the accuracy of data analysis for this case study is around 1 second.

Analyzing the durations that participants looked at different ROIs (LMs) led to some interesting results. The results of data analysis for each group separately are illustrated in figure 4.4 and figure 4.5. Each graph shows the number of selected LMs by an individual user per duration (the spent time on each LM). Figure 4.5 relates to participants of the ET alone method (group 2). These participants used the LMs on the PDA to navigate on the predefined path towards destination without talking. The graphs reveal similar patterns. It appears that participants spent limited time looking at most of the LMs, and there are relatively few LMs at which they spent more time.

Figure 4.4 illustrates the eye patterns of participants in group 1, who used the ET and TA methods. These participants all used the PDA to navigate to the same destination and on the same predefined path. Furthermore, they were asked to mention the proper LMs on the path very briefly. The horizontal axis in figure 4.4 shows the spent time or the 'duration' in which a user looked at a particular LM. The vertical axis shows the number of selected LMs by a user per different 'durations'. All graphs of this group show 2 general trends. The first trend relates to the real worlds' items that were looked at, but not mentioned. This trend (visualized by red bars) is comparable to the pattern observed in group 2. It means limited time was spent on most of the selected LMs. The second trend (visualized by blue bars) relates to LM's which attracted the user (the user looked at them), and mentioned their names as LM. It is observable that almost for all users there are few LM's on which they spent very limited time; on most LM's, they spent between 3 to 9 seconds durations and then numbers of LM's starts to decrease again. The summaries of the 4 graphs for all users in each group are presented in figure 4.6 and figure 4.7. In figure 4.6 the 2 mentioned trends appear more clearly. The major parts of the 2 trend lines represent 2 Gaussian curves with a maximum number of LM's at around 1.5 seconds (they just looked at the LMs), and around 6 seconds (for the looked at and mentioned LMs). Comparing these 2 trends shows that indeed TA in mobile ET affects ET data. When users TA (even with very short phrases), the time spent on ROIs tends to become longer. On the other hand, figure 4.6 shows there are some overlaps between the 2 trends. The overlap starts at the duration of 1 second and increases till it reaches the maximum overlap at around 4 seconds. Then it starts decreasing again. The overlap area represents the extra information which ET can add to the previous field-based methods (Najari, 2009) applied for selecting salient LMs (we may call the overlap area the *'cognitive selected LMs')*. E.g. the graph shows that there are around 20 objects at which the users looked for 4 seconds and then mentioned them as LMs. There are another roughly 20 objects which attracted users' attentions for 4 seconds, but they did not mention their names as LMs. So, it seems logical for selecting salient LMs to investigate the latter 20 objects further to see for instance whether they are common among users. Since the looking durations of the most mentioned LMs here are around 5-7 seconds, it seems logical that the objects with the same or shorter durations which are not mentioned have the highest probability of being potential LMs. Figure 4.6 shows TA affected ET data by making 'looking durations' longer than when users just looked at objects. Hence, objects of the red curve with the shorter durations than the maximum point of the blue curve may also be the potential LMs. Although it was possible to find out what the mentioned and not mentioned LMs were, we did not do it in this analysis, since the type of selected LMs is not relevant for this research. Furthermore drawing conclusion about the type of selected LMs requires more analysis time and participants.

Likewise, figure 4.7 shows that in group 2, the major part of the trend line represents a Gaussian curve with a maximum of around 1 second. Then the trend start to decrease and it reaches zero at around 15 seconds duration.

Figure 4.4: Time spent on each LM for the users of group 1 individually.

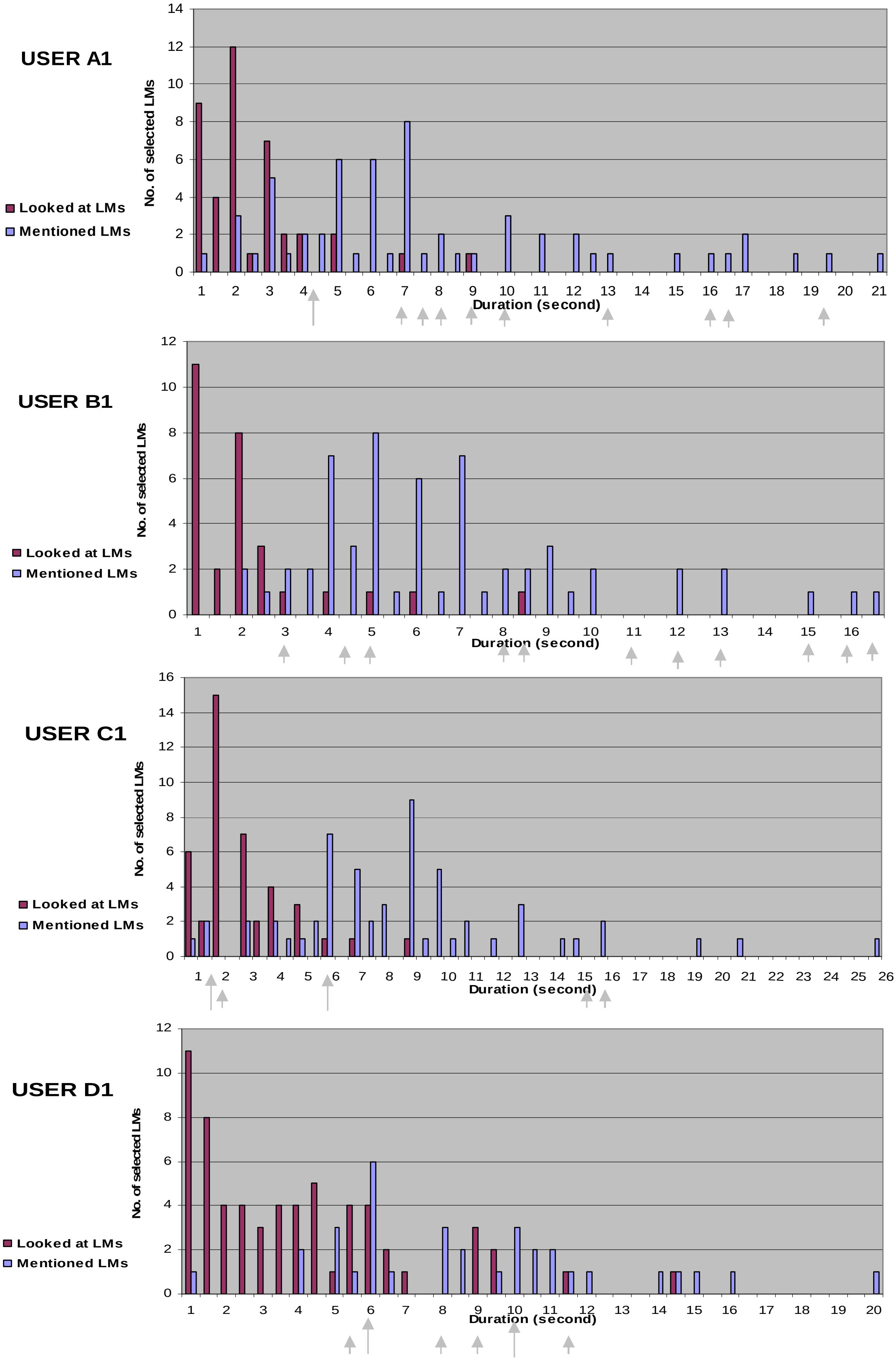


Figure 4.5: Time spent on each LM for the users of group 2 individually.

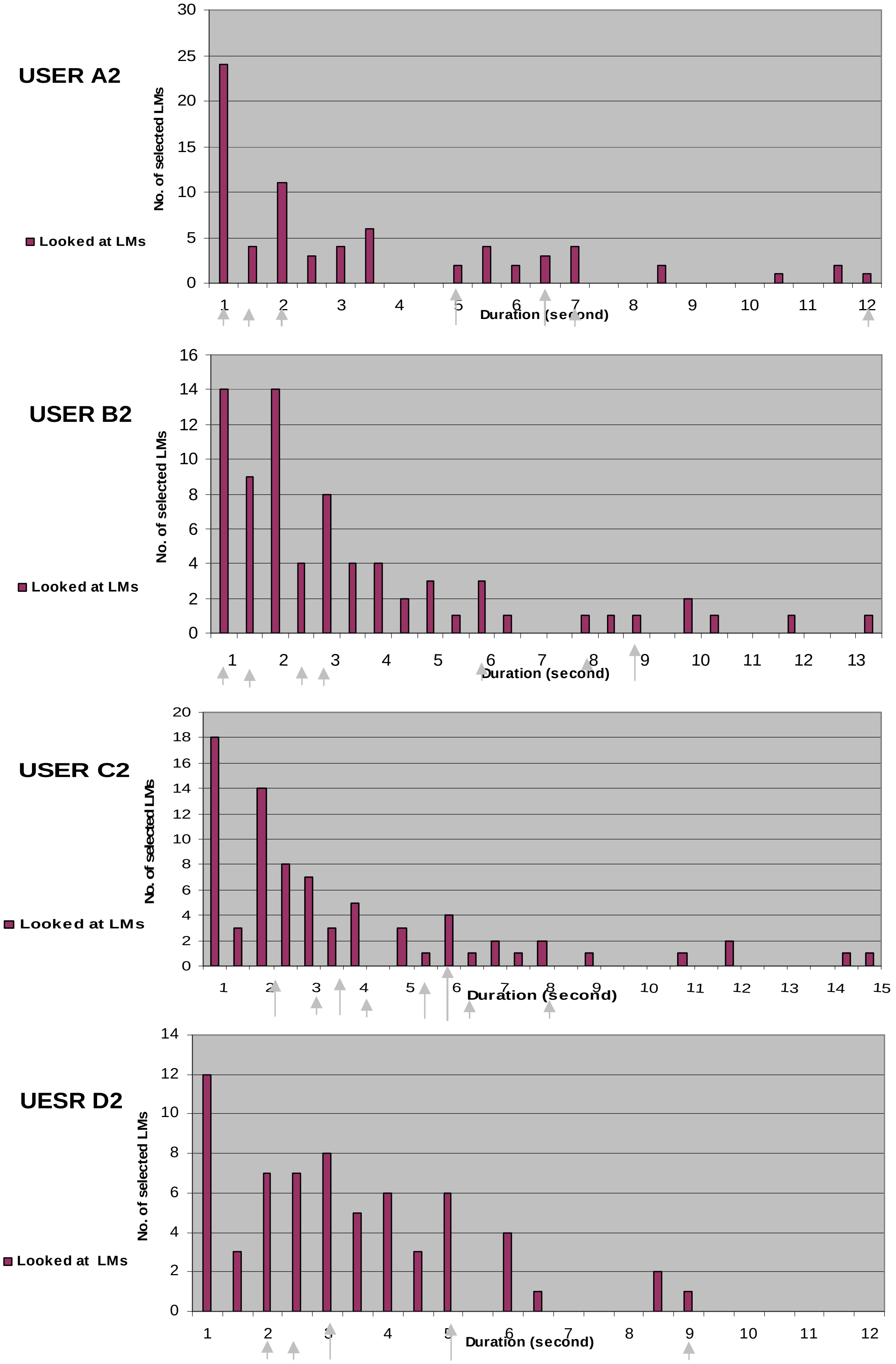


Figure 4.6: Time spent on each LM for all members of group 1.

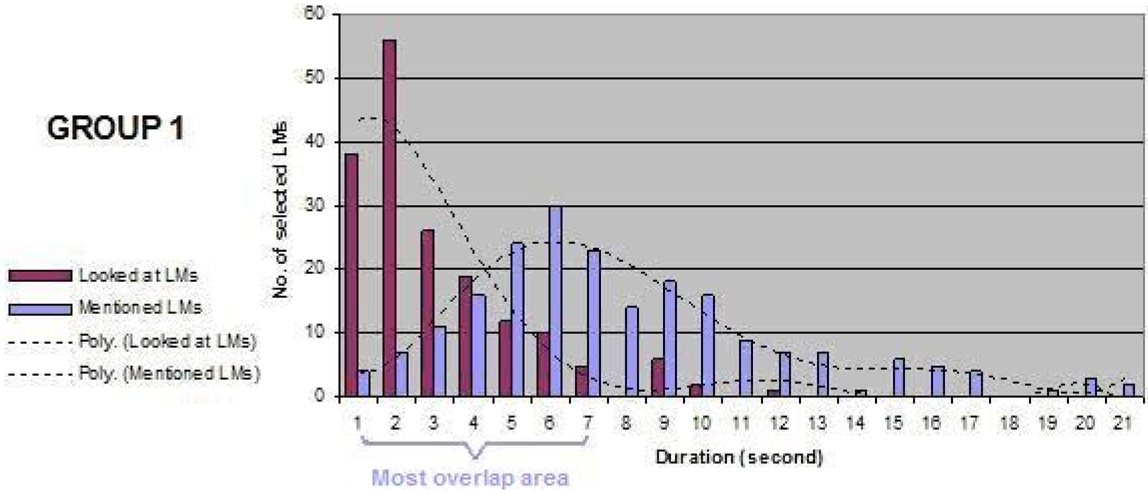
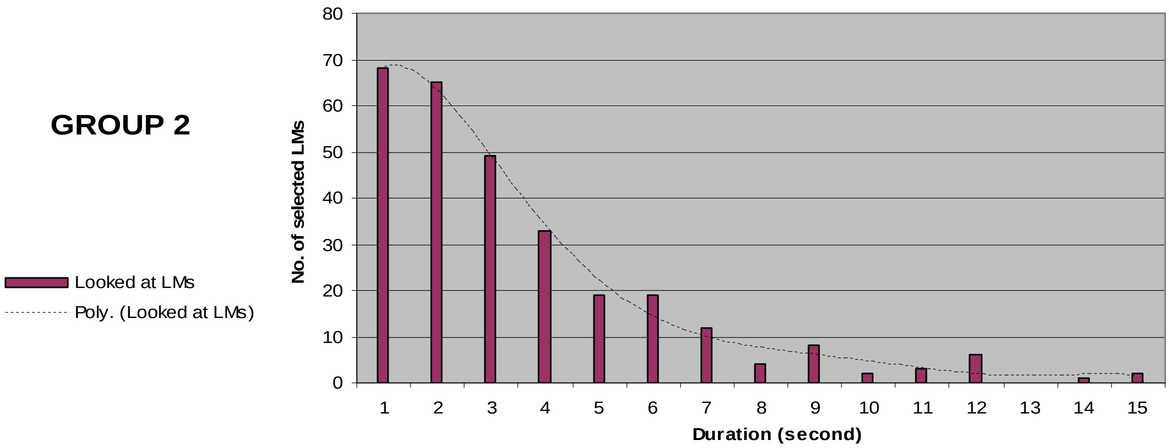


Figure 4.7: Time spent on each LM for all members of group 2.



Some of detected limitations of Mobile Eye after the execution of the test which can affect the results were:

- Although the gaze data were analyzed with an accuracy of around 1 second, there are some factors which reduce data accuracy. For instance, users sometimes get information through peripheral vision without looking straight at an object (LMs). If that is the case, we do not know the exact duration by which the user is attracted to an object. In some cases, it is not clear enough where the user is looking at. Examples of these instances are:

  - When a user turns the pupil too much, the cursor falls on the edges of the video frames.
  - Sometimes when a user is looking at the PDA, due to parallax error, the cursor falls on a real world's LM mistakenly.

In such cases, whenever the error is distinguished no data was recorded for those particular LMs. Still such cases might have reduced the accuracy.

- Another factor which may affect the accuracy of this methodology is that the calculated durations may also depend on the speed with which each individual walks on the path throughout the test. It is possible that a user, who walks faster, has less time to look at a LM while he passes that LM. Due to existing great differences in gaze data between users on an identical task, in order to make a valid comparisons, Hyona (2003) recommends to organize a within-participants design in ET experiments whenever possible. So, maybe it would be wise for improving the accuracy, to design the test within subjects, instead of between subjects.

- One of the drawbacks of the ET technique is that it does not say why a user is looking at a LM. For instance, during the observations the researcher noticed that user D1 in group 1 was looking at one of the shops much longer than at surrounding shops without mentioning that shop (see figure 4.4 - user D1- the red bar at 14.5 seconds). After the test, in answering the researcher's question about the reason of looking at that shop for so long, user D1 said that when he saw that shop, he remembered he required to buy something from that shop and he was trying to remember to come back to buy that after the test. This is one of the shortcomings of using ET for this application. This drawback does not exist in some other user data collection methods which apply audio-video recording of the user in which users could be interrupted by researcher for clarifications, at any moments during the test. e.g. (Delikostidis, 2007).

- The system had problems with users with long eyelashes (and wearing mascara), and it is not possible to apply the system for users with glasses. Usually, it was not sunny during the test. Just in one case when there was sun, (at a low angle in February in the Netherlands) whenever user B2 had the sun in front of her, the eye cursor disappeared. The design of the adjustable headband should be improved to better secure the glasses relative to the head throughout the test.

Analysis of the mental maps

In order to analyse the results of the mental maps and to find out whether there is any correlation between the mental maps and ET data, we first extracted all the LMs illustrated by a user in the mental map, we gave them their related codes (Appendix D), and we made a table out of all the LMs in all the mental maps for the 2 groups. Then we checked the related videos to each mental map and took notes of 2 items: first, we checked whether a LM is 'mentioned' by the user (this was checked for the 1st group). Second, we wrote down the 'spent time' on that LM by that user (this was checked for both groups). The illustrations of the spent time on retrieved LMs from the mental maps for each user are presented in figure 4.4 and 4.5 with some vertical arrows under each graph. The position of an arrow shows the spent time by a user on that LM. The length of the arrow is proportional to the number LMs in related mental map with that particular duration (or the spent time on LMs). Interestingly, the results of the 1st group showed that almost all the LMs of the mental maps were among the 'mentioned LMS' (the arrows are mostly under the blue bars). One reason for this could be that when a user verbalizes an object, it makes him to remember that object better than the objects he does not talk about and just look at them. Table 4.2 provides the summary information of the mental map analysis.

Table 4.2: List of the LMs (codes) retrieved from the mental maps, and the related spent time (digits) on each LM retrieved from the videos.

LM	Group 1 (second)				Group 2 (second)			
	A1	B1	C1	D1	A2	B2	C2	D2
a0	5							
a1			2					
a3	10	13						
b1			2	6			2	
b3	17	16		6			2	
b4	20	17		10			8	9
b5	13			12		9	6	
b7							6	
c0			2					
c1								
c2	9	12		11	5	9	4	
d1					7			
d2	5							
d3		8		6	5	6		
d4		5						
d5								
e2								
e4				11				
e5			6					
e6	8		15				6	
e7						2		
e8					7			
e9					12			2
f4						8		
f8								
f9	8	15						5
g2		3						
g5							6	
h0					7		6	
h9								3
i0							4	
i2							4	
i4					1	1		5
i6					2			
i7		5				3		3
i8					2			
i9	16	9		8		3		
j0								3
j2			16				3	
j3	7	11	6					

The results of table 4.2 (overlaid on the previous information of the 2 groups) are presented in figure 4.8 and 4.9. Figure 4.8 shows that the distribution of the mental maps' LMs is more similar to the distribution of the blue bars (mentioned LMs) compared to the red bars (looked at LMs). Usually, the LMs which a user put in his mental map should be more important for him (since he remembered them among others). So, we may conclude that the 'mentioned LMs' are more important to user than the 'looked at LMs'. Although the 'looked at LMs' provide some user extra information regarding the selection of LMs, maybe the 'mentioned LMs' are the most important for the users.

Figure 4.9 shows the mental maps of the users who used different LMs on the path to navigate without talking about them. Here also the trend of the mental map roughly follows the trend of the 'looked at LMs' (except for the 1 second duration). So, we may conclude here that although a lot of objects are observed briefly (for 1 second) in this method, they are not regarded as LMs by the user. The reason for observing a lot of objects for only 1 second by the user might be that, he is just scanning the environment to find a LM. Based on the mental maps' result, if we apply this method (ET alone) for selecting LMs, maybe we should only look at the objects with more than 1 second duration.

Figure 4.8: Time spent on the LMs which are retrieved from the mental maps for all members of group 1.

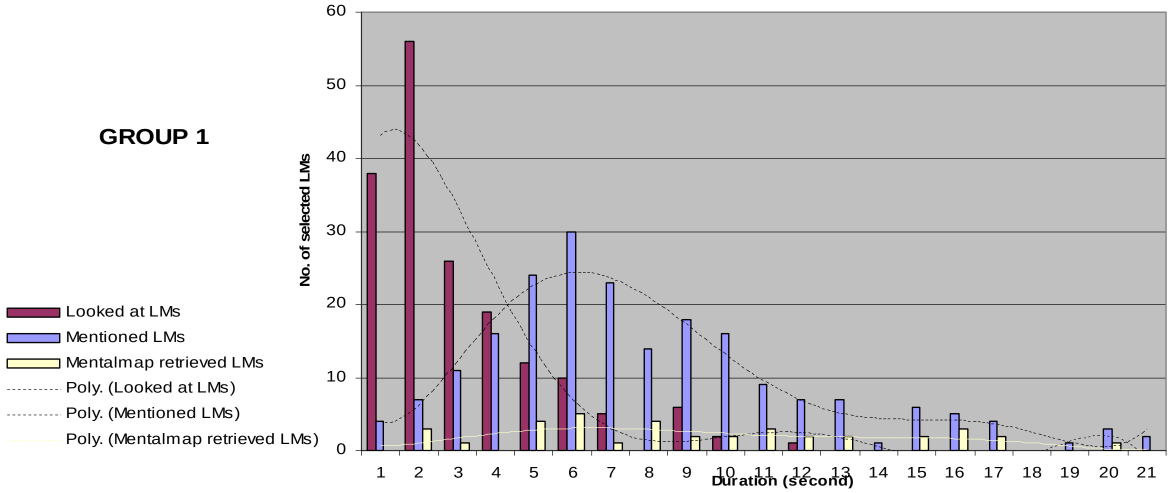
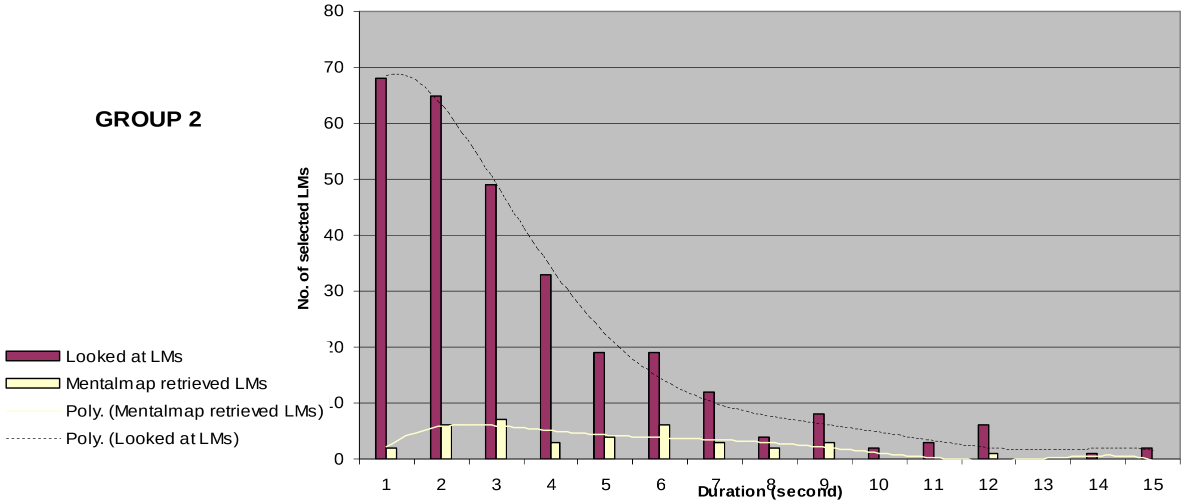


Figure 4.9: Time spent on the LMs which are retrieved from the mental maps for all members of group 2.



4.2.3. Comparing the two methodologies

In this section, the results of applying two different methodologies (ET alone and ET combined with TA) are compared to each other to find out which method is more efficient, effective and satisfactory.

Efficiency is a measure that highly depends on the *time* spent for completing a task. Table 4.3 provides information about the time spent by each user for navigating from the starting point to the ending point to complete the task. It appears that users in group 2 (with the average time of 27 minutes) spent more time than users in group 1 (with the average time 19.5 minutes). Watching the videos shows that the reason for longer test duration for group 2 which only use ET is that they spend more time on PDA to continue the navigation. Hence, according to the results in table 4.3 the method of group 1 in this case study is more efficient than group 2.

Table 4.3: The time spent by each user to complete the navigation task.

Test duration (minutes)	Group 1 ET-TA	Group 2 ET
1st user	18	29
2nd user	18.5	29
3rd user	21.5	28
4th user	21	22.5
Total users	19.5	27

Effectiveness deals with the *informativeness* level of a product (here: a method), and the ease with which a user (here: the experimenter) cando what they intend to do (here: finding users' salient LMs). When we compare the two groups' results, group 1 (ET and TA) appears to be more effective than group 2 (ET alone). Although group 2 shows the same trend as group 1 (red bars), using it by itself is not informative enough for finding the potential LMs. For instance, in figure 4.7 it seems logical to say that when a user looks for a longer time at an object, he probably is attracted more by that object than to objects at which he briefly looked. So we may draw a conclusion that the potential LMs are the ones which belong to longer durations, but we do not have any clue in which duration range users really select (or mention) an object as a LM. On the other hand, the graph of group 1 gives us the information about the time range in which user really selected (and mention) a LM. The overlap area under the two trends lines (the blue bars and the red bars), gives the most probable time range, in which we should look for a potential LM that is not mentioned by user.

Satisfaction refers to *perceptions* and *opinions* of the user (here: experimenter) about the product (here: method). From ther researcher's point of view, for this particular case study, there was no significant difference in satisfaction rates with 2 methods. Both methods had rather similar data collection and analysis issues, and the fact that users applied TA technique in the first method did not affect data collection and analysis complexities. However, we may say the first method provided more satisfactory results. We also assessed the satisfaction factor from users' points of view during a semi-questionnaire after the test. Most users were satisfied with wearing the spectacles (they mentioned they can easily tolerate it, at least for 1 hour); 2 users mentioned they did not like the fact that people look strangely at them. 1 user had physical problem with wearing the glasses. As mentioned, due to the glasses movement during the test, we used extra clips to tighten the headband. The user had long eyelashes, and the fact that the glasses were fastened too tight, was inconvenient for him. The glasses contacted his eyelashes during the test and generated tear drops in his eyes. The recorder was light, and there were no complaints about carrying the waist-bag from participants. Although we had to repeat the calibration, several times for some of the participants, there were no major complaints about the calibration process. The first group was also asked about their impression on doing TA during the test, and they were all satisfied with the technique (ratings were good or very good).

4.3. The second case study: fixed ET

4.3.1. Execution of the test

In order to run the prototype on the eye tracker system, first the required software (Apache server as well as PostgiSQL) were installed. Before the execution of the actual test, a pilot testing was held on 28th of January 2010 during which, the two supervisors carried out the test as participants. However, due to calibration problems, one pilot participant could not continue the test. Based on pilot test experiments, some modifications to the test questions were made by prototype developer. As already mentioned, participants were selected based on a pre-selection questionnaire such that they create 2 homogeneous user groups. There were 4 people in each group. Group 1 comprised of a PhD, a MA and a MSc student and a lecturer. Group 2 included 2 PhD and 2 MSc students. The average age of participants was 32 and they were from different nationalities. The test was held on the 1st of February in the ET lab at UT. It started at 9 AM and finished at 6 PM. Each participant did the test individually. The average time spent by a test participant was about 1 hour. The ET data of each subject was recorded via the software immediately after finishing each test and before the arrival of the next subject. Unfortunately due to saving problem of the software, we lost the data for the 3rd and 5th subjects. The remaining subjects comprise 6 participants: 3 participants for the first method (ET and TA) and 3 participants for the second method (ET alone). The summary of background information for the 6 subjects is compiled in table 4.4; it can be noticed that the groups are relatively homogeneous.

Table 4.4: Summary of background information for each user group.

Familiarity rate with: (percentage)	Group 1 ET-TA	Group 2 ET
trajectory data	80	80
icebergs	45	55
interactive map	85	80
animation	80	80
websites	85	80

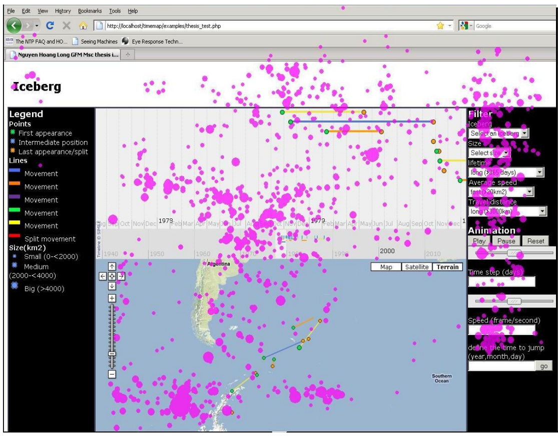
4.3.2. Analyzing procedure and results

The extracted data files from faceLAB software included: 1) txt files of fixation with their attributes including a code for each fixation, x, y, starting time and ending time. 2) avi files including the stimuli with the overlaid eye cursor and the synchronized user's voice. The analysis involved watching the video files as well as getting some statistical outputs from the software. Since this user testing was a joint case study of the researcher and the developer of the prototype, the data were looked at and analyzed from 2 different points of view. One of the objectives which is not much relevant to this research is to assess usability of the prototype itself as design solution for visual exploration of iceberg data. This section describes the characteristics of the prototype which we used for the test and some of the revealed prototype issues by the ET technique. Another objective in designing the case study which is more relevant to this research is to compare the two different methodologies (ET alone and ET combined with TA) in this geo-application to find information regarding their efficiency, effectiveness and satisfaction rate in obtaining user data. This comparison is provided in the next section.

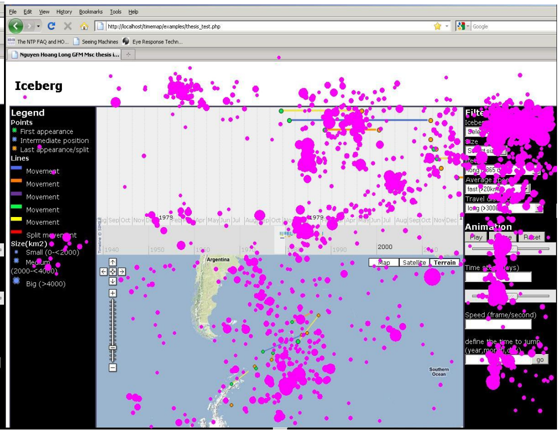
Regarding users' extra information by the ET technique, due to the speed issue in the running application no visual or statistical patterns were observed between the 2 groups. The speed problem was caused by the limited capabilities of the hardware in running the stimulus. Due to this issue, the software could not run the stimulus at an even pace for different questions and different participants during the tests. Some of the outputs like distribution of fixations and the times spent in 4 defined look zones , for each individual users are provided in figure 4.10 and 4.11. The same results are provided for each individual group in figure 4.12 and 4.13. Finally, the same results for all users are provided in figure 4.14. Due to inaccuracy because of the speed issue, any comment on these visualizations is not definite. Nevertheless, looking at figures 4.10 and 4.11 we realize that look zones 2, 3 and 4 (timeline, functions and map view) are the main used areas by all users. None of the users looked much at the legend. This shows that during pre-test practice they got familiar with the legend. It also shows that the legend is simple and easy to remember. So, maybe it is possible to present more detailed information about icebergs via the legend to the user. Figure 4.10 shows that there are many fixations on the upper white area (outside the application). Watching the videos revealed that they are mostly because of the system's speed problem which made the users look around the page, and also partly because of error in calculating fixation locations by the software. Figures 4.12 and 4.13 show that the 2 groups paid attention to the 3 look zones (2, 3 and 4) more or less evenly. Most of the test questions were organized in a way that they could be answered by using either the timeline or the map view. Before the test, we expected that users would spend most of their attention on the timeline and on the map view. It means if a user chose randomly the timeline or the map view to answer a particular question, we can statistically expect that users use the timeline and map view almost evenly. If the concentration of fixations in one of these 2 look zones were significantly different from the other one, we could think that one the look zones had probably some problems to be used, or it was more difficult to understand. Figure 4.14 shows that the amount of time (and fixations) which users spend on the map view is roughly similar with the time spent on the timeline. The results of the ET data are more or less similar to our expectation; it means the timeline and map view both work rather well.

Finally, the amount of time (and fixations) that participants spend on lookzone 3 (functions) is also roughly similar with the timeline (or map view). However, the expectation was that users would apply the functions more efficiently. Normally when users want to perform a filtering or animation function, they should look at the function zone and then click on the required function; this process should go very fast. According to Poole and Ball (Poole and Ball, 2005) if participants pay attention at one item longer than expected, it indicates that this item may lack meaningfulness and probably needs redesign. Based on the result, we think that the interface functions may be difficult for users to understand. Watching videos shows that there are some problems with the functions which sometimes confuse the user. For instance, some filtering options (e.g. lifetime, average speed, travel distance) can only be applied for the overview data set. Another example is that the combo-boxes in the filtering functions did not refresh by themselves. For instance, if the combo-box of lifetime already shows the value 'long', and if we want to filter the icebergs that have a long lifetime, we should change the combo box to another value (e.g. medium) and then choose a long value again.

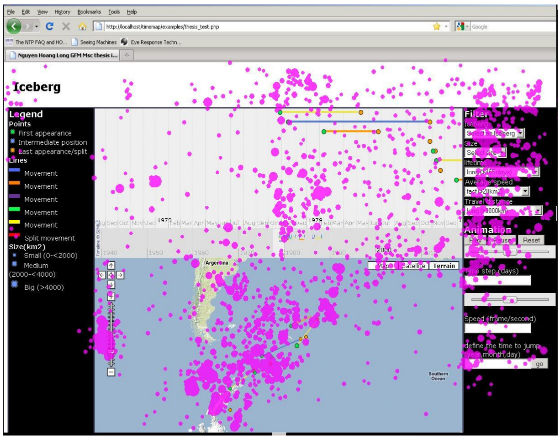
Figure 4.10: Fixations locations (circles) and their durations (circle's diameter) for each test after shift correction. Due to a speed problem in running the application on the available hardware, no different visual or statistical patterns were observed between the 2 groups. There are many fixations on the upper area (outside the application), mostly because of speed problem, which made user look around while waiting for a system's response.



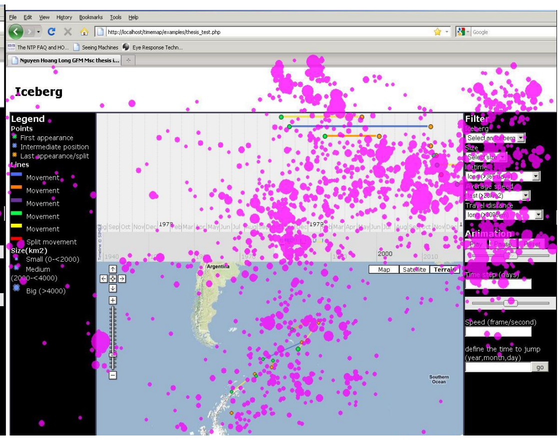
Group1 _User A1



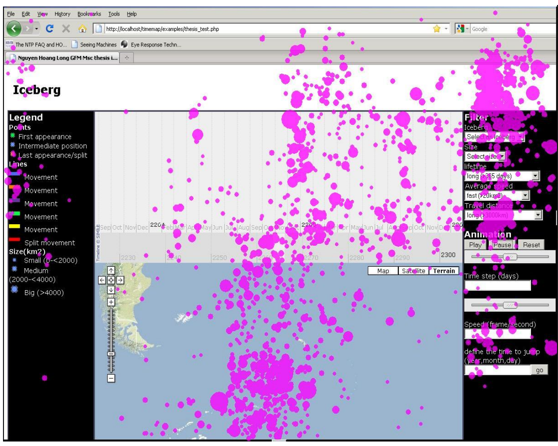
Group2 _User A2



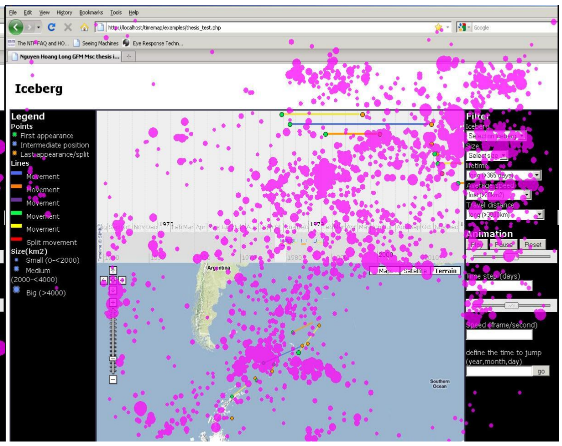
Group1 _User B1



Group2 _User B2

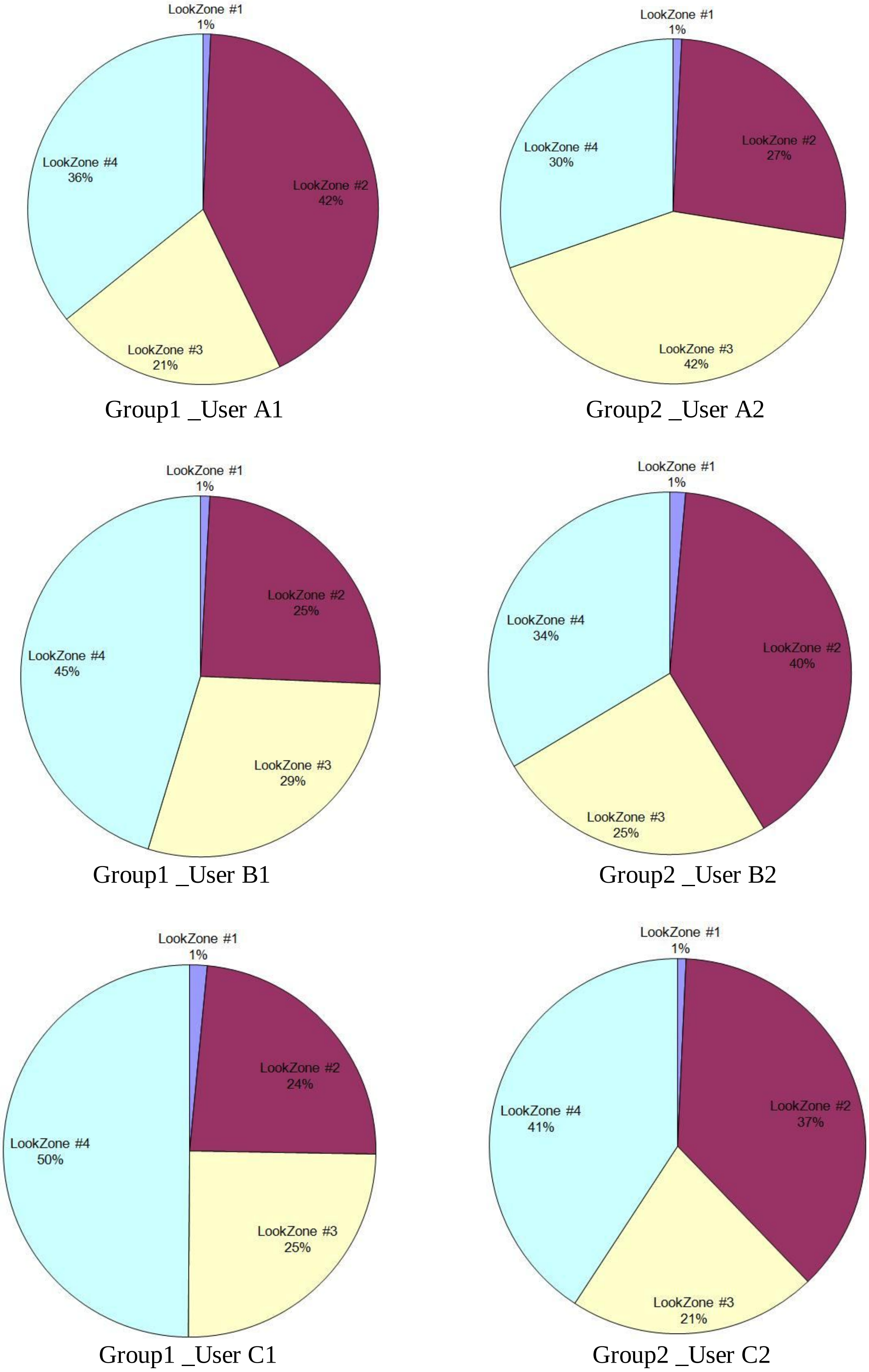


Group1 _User C1



Group2 _User C2

Figure 4.11: Percent time in the 4 defined lookzones: lookzone1= legend, lookzone2=timeline, lookzone3=functions, lookzone4=map.

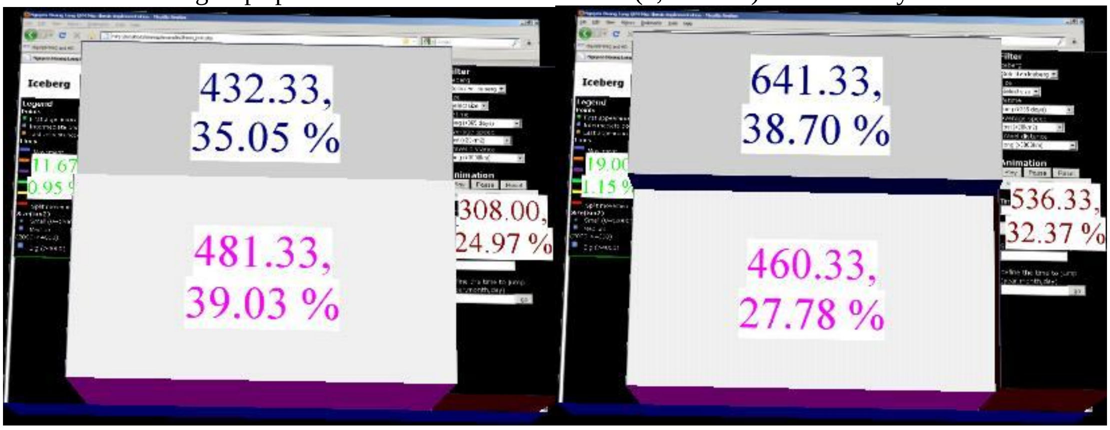


Group1 _User A1 Group2 _User A2

Group1 _User B1 Group2 _User B2

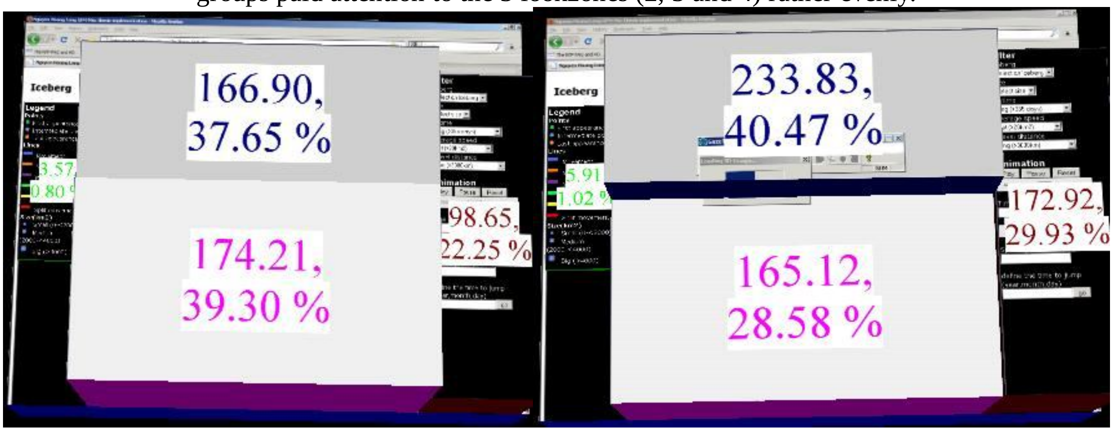
Group1 _User C1 Group2 _User C2

Figure 4.12: Average of total fixations in different lookzones for each group individually. It appears that 2 groups paid attention to the 3 lookzones (2, 3 and 4) rather evenly.



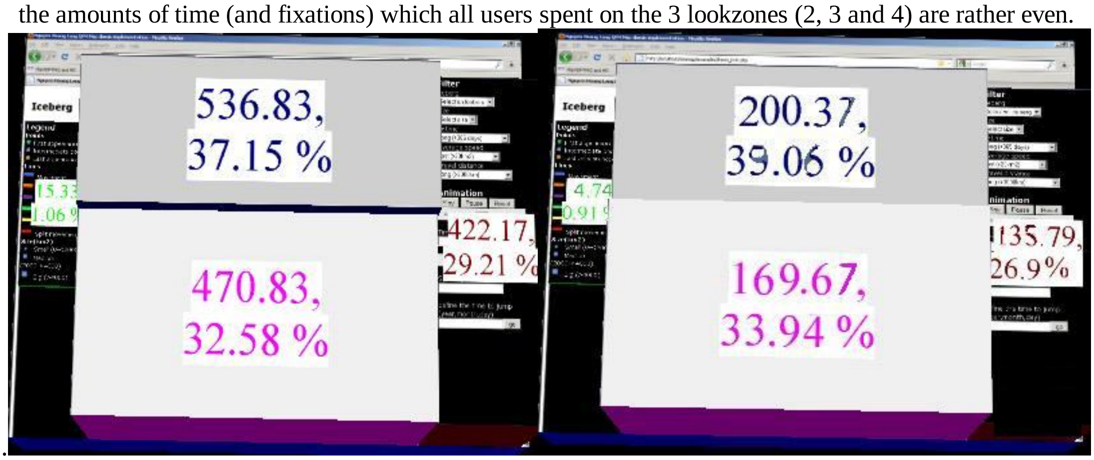
Group1 Group2

Figure 4.13: Average of total spent time in different lookzones for each group individually. It appears that 2 groups paid attention to the 3 lookzones (2, 3 and 4) rather evenly.



Group1 Group2

Figure 4.14: (a) Average of total fixations and, (b) Average of total spent time in lookzones for all users. It shows that the amounts of time (and fixations) which all users spent on the 3 lookzones (2, 3 and 4) are rather even.



(a) All users (b) All users

4.3.3. Comparing the two methodologies

Efficiency relates to the spent *time* for solving a task. As mentioned in section 3.3.2.2, due to a hardware problem, the system was randomly slow, and the software could not run the stimulus at an even pace for different users as well as questions. Analyzing the videos shows that a longer spent time does not necessarily mean the user is busy with finding the answer to a question. Sometimes the user is doing nothing and just waiting for a system's response to take the next action. Because of this issue, no conclusion regarding efficiency can be derived in comparing the 2 methods. Also, due to the speed problem assessing a hypothesis like 'fixation durations in ET and TA are longer than in ET' (because 'when a user is talking, he may look for a longer time on an object.') is not possible. Assessment of this hypothesis could provide the answer to one of the research questions which is 'how does TA affect ET?'

Effectiveness as mentioned, deals with the assessment of how much a product (here: a method) is *informative*, and the ease with which a user (here: the experimenter) does what he intends (here: finding the prototype's problems). The statistical analysis of ET data (e.g. fixation calculations, durations, etc) is usually done by ET analyzing software automatically. Statistical analysis can also be done manually or via statistical packages. In either case, there is no difference in analyzing ET data to be with or without TA data in providing more user information. However, as it comes to animated data (eye cursor over stimulus), ET data combined with TA are more revealing than ET alone. In replaying the videos of the tests of the first group (ET and TA), we see that while the user tries to find an answer we have a step by step explanation for each eye movement. For instance, when he says I want to go to this specific date, and at the same time he is looking at the animation and jump-to-time function alternatively, we know that he cannot understand which function works more properly for this question. For the second group (ET), although looking at the eye cursor gives some clues about what the user is trying to do, it is not clear enough how finds the answer.

Satisfaction as mentioned, is about user's *judgments* (here: experimenter) and opinions about the applied product (here: method). From the researcher's point of view, since ET and TA make the analysis simpler, it is a more satisfactory method than ET, provided that TA does not affect much the eye movement patterns. Since we could not compare eye patterns on an even pace, we do not know how much eye movement patterns are really different in ET compared to ET and TA. Hence, a more accurate comment about this is not possible. Satisfaction rate was also assessed from the users' point of view via the semi-questionnaire. It was revealed that all 8 users were satisfied with the using the ET technique (they all rated applying the technique as good or very good). They found the method interesting. They suggested that they were comfortable during the test, and said the fact that they had restricted movements during the test did not bother them. The only complain was from the Dutch lecturer about not having a comfortable position during the test, probably because he was taller than other users, and we could not reduce the chair's height more to make him more comfortable. The first group was also asked about TA experience. They were all satisfied with applying TA (ratings were good or very good). They suggested that talking about what they thought did not bother them during task completion.

4.4. Conclusion

Chapter 4 described the actual test execution for the 2 earlier defined case studies with the mobile and stationed eye trackers. Then data analyses for the 2 case studies based on the objectives of the research and considering the existing limitations regarding the analysis procedure are described. Due to these limitations (especially for stationed system) analyses could not done completely. Finally, the results of data analyses of the 2 different methodologies were compared for each system separately. In the 1st case study, the results showed that ET revealed some extra information (or LMs) which was unconsciously selected by users, although they were not mentioned. Due to some technical problems, data analyses in the 2nd case study were not complete. However, the results showed much more intensity of gaze data than expected in the functions zone which made us think that the interface functions create some difficulties for users. Regarding the comparison between different data collection methods (ET, and ET with TA), in both case studies ET with TA was recognized as the preferred method. The next chapter describes a summary of whole research, conclusions and further required enhancements which were not applicable during this research. Finally, some recommendations for the usability of ET systems in the geo-domains will be suggested.

5. Conclusions and recommendations

5.1. Summary and conclusions

The present thesis investigated the usability of ET as a user research technique in geo-information processing and dissemination. For a long time in the past, there was no attention towards users as part of investigating developed technologies, and it was not clear whether created products were indeed usable. Usability, however, proved to be an important part of investigating new technologies in different fields including geo-informatics. Today, there is gradually more attention towards testing the usability of different products as well as methods and techniques for doing use and user research. ET is one of these techniques which tracks and records eye movements of a person via an eye tracker, while they see, for instance, an interface. It usually works by reflecting IR light onto an eye, recording the reflection, and calculating point-of-regard by a geometrical model. ET has been applied to some extent in usability researches in geo-informatics, and there are reports of ET enhancing conventional usability testing by revealing hidden cognitive processes of human beings. However, there are some uncertainties about its usability such as the availability, costly performance, intrusiveness, robustness, handling data and analysis complexity with no warrant of getting extra user information, etc. There is also the issue of combining the ET technique with other techniques of usability such as TA. Due to the sensitive nature of eye movements, TA can change gaze durations, since users need time to verbalize ongoing cognitive processes. On the other hand, some researchers proved that concurrent use of TA with ET could lead to valuable results. Assessments of these issues were the main motivations behind this research.

In chapter 2, we discussed all general characteristics of the ET technique. First, concepts of usability testing and its different methods and techniques applied for evaluation of different (geo-) applications were investigated. Then the ET technique, some history, and some applications (including in the geo-domains) were briefly described. A typical ET study usually consists of different stages of data collection, data clustering, data analysis and data interpretation. The rest of chapter 2 was devoted to review and assessment of various existing methods applied in different stages of an ET study. In the end the main benefits, drawbacks and limitation of ET are identified.

Considering the cost-benefit perspective, in Chapter 3 we tried to apply this technique in two case studies to observe its usability, potentials and limitations in geo-domains in practice. For this purpose and in order to get a general overview regarding the usability of ET with different systems, we designed 2 case studies with 2 main types of ET systems (mobile and stationed) to investigate the selection and use of different methodologies for the usability testing of 2 geo-applications. Also, in order to assess the effects of conventional techniques (TA) on ET and find the most resultful combinations of these techniques, we applied 2 user groups with different methods of usability in each case study. These combinations were the use of ET alone and ET with TA (together with mental map and questionnaire) for the 1st case study, and using ET alone and ET with TA (together with questionnaire) for the 2nd case study. During data collection and analysis of these 2 case studies, some issues related to clustering, analyzing and interpreting the data were discovered and discussed in chapters 3 and 4.

The main results of this research are summarized by answering the respective research questions, framed in the first chapter, as follows:

- *General questions:*

  1. *What is ET? What is it applied for? What are possible applications and capabilities in different domains? What are the related experiences in the geo-domain what were the comments upon the usability of the resulted data?*

The answers to these general questions are provided in chapter 2. Sections 2.2 and 2.3 define the ET user research technique and its position in usability testing. Possible applications in different domains and in the geo-domain were described in sections 2.4 and 2.5. The known obtained experiences in the geo-domain show that ET can enhance conventional methods by giving new insights into the usability issues.

2. *What are the advantages, disadvantages and current pitfalls of the technique? Are the disadvantages manageable?*

Section 2.7 provides detailed information regarding the advantages, disadvantages and pitfalls of ET found in literature. During the 2 case studies we experienced some of these (dis)advantages regarding data collection, analysis and interpretation. All the weak points and encountered issues as well as the potentials and obtained results for the 2 case studies are described in details in chapter 3, 4 and 5.

3. *In which stages of the UCD process of the geo-information domain can the technique be applied?*

According to literature, ET is often used in the prototype and evaluation stages (2^{nd} and 3^{rd} ovals in UCD process) of a (geo-) application. During the 2 case studies we experienced the applicability of ET in the 1^{st} stage (for the 1^{st} case study), and at the 2^{nd} stage (for the 2^{nd} case study) of UCD. The results of the 1^{st} case study confirmed the possibility of applying ET in the 1^{st} (requirements analysis) stage of UCD in designing a prototype for a pedestrian navigation system. The 1^{st} case study proved that extra user info can be obtained via ET. As it comes to the 2^{nd} case study due to the hardware problems of the system for analyzing ET data, the analysis was not completely done. However, analysis of the ET data confirmed that there were usability problems with the functions of the prototype regarding problems related to the functions. If the hardware problem could have been solved, the 2^{nd} case study would most likely lead to some extra user information.

4. *What other factors can affect (improve or bias) the results? Does a learning effect, environmental condition, test duration or other factor modify the test results?*

Depending on the application and the type of applied device, there are some factors which can affect the results. During the 1^{st} case study we experienced some of these factors. For instance, we had to tighten the glasses firmly to prevent its movement during the test which was not so convenient for users. Although users did not complain about that, during the semi-questionnaire almost all users mentioned they can tolerate the glasses for 1 hour. This will most likely affect the results for longer tests. Weather conditions can also affect ET results. For instance it was noticed that when it was very cold, that users made more mistakes in navigation and they could not concentrate properly. Also, the lighting condition was a factor which affected data analysis. Due to this we had to organize the test in brighter hours of the day. As it comes to the 2^{nd} case study, we observed that a hardware problem led to the software problem in running the application. Also, during saving the data of some users, due to a software problem we lost the data (this, also happened in the 1^{st} case study). Finally, the familiarity rate of the users with the study area (the 1^{st} case study), or the prototype (the 2^{nd} case study) are other factors which can bias the outcomes. Due to the selection of users with roughly the same background information, we ignored this bias for our tests.

- <u>*Data collection questions*</u>

1. *What kinds of raw data are collected by an ET system?*

2. *How is the ET system applied? What is a typical test procedure? What does the test person/experimenter typically do during the test? Does the device type (mobile/fix) effect the executing procedure?*
3. *What kind of different components do ET systems include? Do the equipments differ for different systems (mobile/fix)? What do all components do?*
4. *What are the different capabilities of the two (mobile/fix) systems? Can mobile systems be applied for document based and computer screen studies as well?*
5. *What is the general workflow of ET systems in order to record the track of the eyes?*

Section 2.6.1 provides the answers to all of these questions in details. The types of raw data, different components of different systems and their functions, different capabilities of the 2 main eye trackers, and a comparison between them and remaining questions were answered in this section. Also, the different stages of a typical ET test procedure are explained in section 2.6.1.4.

6. *Should ET be combined with other usability techniques? Which method (alone/combined) is preferred as a user research technique? How should ET be combined with other usability techniques?*

Our experiments showed that combination of ET with TA (if applied correctly) provides extra user information: in the 1st case study, the results of the 1st group showed we can apply TA with mobile ET, provided that it is applied correctly, depending on the application (e.g. using short phrases for this application). As we saw in the 1st case study, TA affected ET data. However, the outcome showed that TA had roughly the same effect on gaze data for different users (the blue Gaussian curve in figure 4.6). This makes it possible to compare the results of different users who made use of ET combined with TA. We can simply say that TA pushed the original normal distribution graph towards longer times that participants looked at LMs. For the 2nd group (although the same pattern as the 1st group is observed), we do not have any clue to distinguish in which duration range we may identify a 'looked at' object as a LM. In fact, TA gave us some clue to obtain extra user information regarding the test objective.

Also, in the 2nd case study when we used ET combined with TA, data analysis (videos) was easier than when we analyzed ET data alone.

The way ET should be combined with TA depends on application and test objectives. For instance in the 1st case study, in order to collect user data (which were the mentioned LMs and the looked at LMs) on an even pace, we asked users to use short phrases. However, in the 2nd case study for analyzing gaze-over laid videos, there was no such limitation.

7. *What is the use and the effectiveness of using the ET technique compared to other comparable user techniques like audio-video recording? And which method is preferred?*

Choosing a proper method for usability testing also depends on the application and test objectives. The results of the 1st case study showed that ET (for this particular application which is selecting salient LMs) is preferred over some other already applied methods, e.g. (Najari, 2009), since it provides information about 'human cognitive aspects' in selecting salient LMs which are closer to reality than the 'logical reasoning' (in which user were asked to say why they selected an object as a LM). The main shortcoming of ET is that we do not know why users choose a particular LM. It was not possible to ask users about their motivations in ET during this test. The reason was that users usually looked at a particular LM several times, even after they mentioned a LM they often turned to look at it again several times. Any communication to the user could interrupt the duration and the number of times a user may spend to look at a LM. The researcher's proposed solution for this shortcoming is that we use RTA (retrospective thinking aloud) added to concurrent TA, and ask the user about their motivation on selected (mentioned or longer looked at) LMs immediately after the test. This probably requires using shorter routes in order to help the user to remember what he did throughout the test.

8. *Which preparatory activities, assistance, further equipments or other material are required? What are the costs involved (in terms of manpower, equipment, time)?*

ET is a very sensitive usability technique. Before implementing an ET test we must be sure about any details regarding the test, since any trivial issue may affect the results. E.g. if possible the pilot testing should be done with more than 1 or 2 users (most of the issues could have been revealed earlier if we had time for more pilot testing). The questions (in the 2nd case study) should be placed so that it does not require head movements. To prevent wasting time of the user and researcher, there should be more time space between the sessions of 2 subsequent users than in a usual usability testing (due to potential problems like calibration, or saving data), specially if the lab is rented.

Regarding manpower for both systems, it was noticed that if everything is scheduled, one person (researcher) will be enough to manage the test and no assistance is required.

- *Data clustering and analysis questions*

1. *What are the impediments of clustering data? What (visual/automatic) clustering and analyzing tools are currently available? Can available softwares meet the needs of clustering and analyzing data in the geo-application domain? Do the softwares differ regarding the system types (mobile/fix)?*

The type of clustering algorithm can have dramatic effects on analyses and interpretation of the data. There are many available algorithms and the outputs of different clustering methods are not identical. In practice, usually only a few of many available clustering algorithms are applied (which are briefly described in section 2.6.2.2), however, the same algorithm will produce different outputs by changing its parameters. According to literature and our experience, there is a lack of open processing methods regarding clustering algorithms in available ET software. If existing, such information will provide options for users to choose and will give them outlooks about what results they can expect and why. This leads to more accurate data analyses and interpretation. Due to time limitations of this research, no further research regarding proper algorithms for different applications was done.

Presently, there are many clustering and analysis software tools available from different developers and ET companies, that each provides different features and visualization tools to support one or more ET systems. Also, different software is used for mobile and fixed systems, however, some software can support both. For instance, the gaze tracer which we used for the 2nd case study could also handle the data from the 1ST case study. The classifications of these tools with their different specifications are provided in literature e.g. (Špakov, 2008). Still, the available tools are not completely able to respond to the requirements for different (geo-) applications. During the 1st case study we experienced some of these limitations:

Due to limitations of this research in the 1st case study, we analyzed the data manually. To speed up the analysis procedure in Mobile ET, it is necessary to use an ET software which enables us to define ROIs, and then the software calculates the gaze data in each ROIs. The relevant software (e.g. gazemap) for such applications use some feature extraction object processing algorithms to map fixation points to user defined ROIs. However, in communicating with ET companies like ASL, we found out that the present software is not suitable for our application. This software is still in early stages and works properly for more categorized ROIs like the shelves of a supermarket and not for selecting LMs for a walking user (figure 2.10). It seems the ET software for applications like ours require to be developed further. Since we did not apply the software for our analysis, we cannot comment on this further.

2. *What are the defined ET metrics for analyzing the data? Which metrics (fixation, saccade, AOI and etc) are proper to be analyzed in different geo-applications?*

There are many available ET metrics (and several non-gaze related metrics like pupil diameter) used in literature. Different attempts in ET studies have led to a heterogeneity regarding the terminologies and definitions of ET measurements. A thorough discussion regarding ET metrics and the ones typically applied in the geo-domains are provided in section 2.6.2.1. Although some of these metrics are much more common for data analysis, the types of applied metrics depend on the application and required results.

3. *What other measures (video, sound, mouse click, etc) can be combined with ET metrics to improve analyzing the data via available software? Can we use combinations of these measures via available software?*

Some recent ET systems (e.g. Tobii) provide the opportunity to integrate gaze data and other data from the operating system like key strokes, mouse clicks, document scrolling, window sizing and web page URL recording, along with other user data like voice, facial expression and behaviour through audio and video recording during and even after the test. All these data along with the test set-up information reside in a built-in database that could be used later to extract relevant data for some particular processing. Using combinations of these measurements (off-line or real-time) provides a deeper insight into users motivations while carrying out a task and raise the quality of data analysis. The software we applied in the case study (gaze tracker) was able to integrate gaze data with user's voice, key strokes, mouse clicks, document scrolling, and web page URL recording. However, due to hardware problem, the obtained data quality was low. Integration of facial expression via video recording was not possible with this software.

- *Data interpretation questions*

1. *What kind of results can be obtained with this technique in order to improve the usability of geo-information applications?*

The types of obtained results of an ET study (which often are confirmatory or extra user information), depending on application vary. ET can reveal valuable information about hidden cognitive processes of human behaviors by recognizing where, when and what exactly users have looked at, which cannot be reached by conventional techniques. In the 1st case study we experienced the usability of ET for selecting salient LMs in the 'requirements analysis phase of a geo-application (pedestrian navigation system) which is supposed to be designed later. The result showed that ET (like other already applied methods) was successful in finding user selected LMs. Furthermore, ET could reveal some extra information (or LMs) which was unconsciously selected by users, although they were not mentioned. Although possible, further analysis regarding the 'types' of LMs was not done during this research.

ET can assure the design quality of a prototype, by revealing relative intensities of a user's attention to various parts of a prototype. During the 2nd case study we experienced the usability of ET in the 'produce design solutions' phase of a geo-application (prototype for visualization of icebergs), and assessed its design solution for visual exploration of iceberg data. Although the analyses were not complete, the results showed much more intensity of gaze data than expected in functions zone. This makes us think that the interface functions create some difficulties and that users do not easily understand them, so they probably need to be redesigned.

5.2. Recommendations

In this section, based on available resources as well as our experiences, we propose some usability recommendations for the applied systems and the resulting data and also for the method in the geo-domains, in general.

5.2.1. Mobile ET systems

- Due to the parallax error, using Mobile ET for pedestrian navigation tasks with the present technology seems impossible. Due to this issue, it is recommended that ET companies find better solutions for improving the accuracy for near and distant foci in Mobile eye trackers and for navigation applications (which is probably the most important mobile geo-application).

- Since the objective of this case study was to assess the usability of ET by comparing different methodologies (and the effects of TA on ET), and also because of the time constraints we did not assess the types of LMs users selected by each method. This analysis may lead to finding some trends about the types of potential selected LMs by different users, especially in the 'overlap area' (figure 4.6).

- During analysis there was some trend noticed between the mentioned LMs by different users and the number of fixations (or looking at the same object again), which also because of time limitation, was not included in analysis. It means whenever users mentioned the name of a LM, they usually looked at that object, several times (before, during and after mentioning its name). This did not necessarily happen to the 'just looked at' objects. This correlation may be used to confirm the potential LMs.

- Since the calibration may decrease after some time for different reasons, it is always necessary to have a laptop with installed calibration software in order to do the calibration at the starting point and in the field.

- We executed the test under no time constraint, and in an environment that was rather familiar to all users. The test may be repeated under different conditions like under time pressure (in which users are forced to complete the task as quickly as possible), or in an environment which is totally new to the users, to see whether the results would be the same or not, and how these factors affect ET data.

- Due to system's problems and time constraints, each method was completed for 4 users. We also used ITC GFM students (which all have a high level of familiarity with maps, navigation task and LMs) as users for executing this test. It would be wise if the test can be repeated with more participants including other usual end users of a pedestrian navigation system.

- As mentioned in section 4.2.2, we can extract the potential LMs which are not mentioned, from the overlap area with the highest probability. However, determining the exact boundary between an object to be regarded as a user's selected LM or not, requires more experiments and investigations.

- It is better to always have a back up of evi file (which keeps the calibration information) for each user. If something happens to evi files, related avi data files cannot be used.

5.2.2. Stationed ET systems

- ET is a proper technique to add extra user information to usability testing if it is applied properly and with proper devices. Regarding the quantitative data of ET which relates to measurement of spent time on different objects; system's speed is a very important factor in obtaining reasonable data. The faceLAB eye tracker we applied for the test was not new and probably applied before for many other usability tests. During an introduction to the system by the owners, they did not say anything about the system's slowness in running an application. It seems the system had better speed and accuracy for previous tests which were shorter and used simpler applications like slides, images, texts, etc. Eye trackers are developing day by day, both regarding hardware and software, using advanced technology. However, the new applications which require usability testing are also becoming more complicated every day and they require more speed and accuracy regarding ET data collection and analysis software. The prototype we used was developed for an MSc work, and we used just a subset of the real data for the test. Fully developed applications will definitely need a much better ET hardware and software, especially in the new geo-applications, since they mostly deal with using web, interactive maps, animations and complicated functions.

- The minimum of 3 participants for an ET test, even if they produce acceptable data may not reveal a pattern that can be compared to another groups or lead to a conclusion. It is better to use more participants. The main factor that affected the results of our test was the slowness of the system; however, the number of users can also influence finding a pattern in the 2 user groups.

- The applied ET software (faceLAB) for this case study did not provide clear information regarding the applied algorithms for data clustering. According to literature, other ET software also has this limitation. For our case study the analysis was not done completely, however, having more knowledge regarding the applied clustering algorithms and their effects on resulting data can provide options to choose, which leads to more accurate results for different applications.

References

ALAÇAM, Ö. & DALC , M. (2009) A Usability Study of WebMaps with Eye Tracking Tool: The Effects of Iconic Representation of Information. Springer.

BACH-Y-RITA, P., COLLINS, C. C. & HYDE, J. E. (1971) The Control of Eye Movements, New York, Academic Press.

BALL, L. J., EGER, N., STEVENS, R. & DODD, J. (2006) Applying the PEEP method in usability testing. Interfaces, 67. 15-19.

BARNUM, C. & DRAGGA, S. (2001) Usability testing and research, Allyn & Bacon, Inc. Needham Heights, MA, USA.

BATES, R., DONEGAN, M., ISTANCE, H., HANSEN, J. & RÄIHÄ, K. (2007) Introducing COGAIN: communication by gaze interaction. Universal Access in the Information Society, 6**,** 159-166.

BOJKO, A. & SCHUMACHER, R. M. (2008) Eye Tracking and Usability Testing in Form Layout Evaluation, User Centric, Inc.

BRENNER, C. & ELIAS, B. (2003) Extracting landmarks for car navigation systems using existing GIS databases and laser scanning. International archives of photogrammetry remote sensing and spatial information sciences, 34**,** 131-138.

BRODERSEN, L., ANDERSON, H. K. & WEBER, S. (2002) Applying eye-movement tracking for the study of map perception and map design. Publications Series 4. Copenhagen, National Survey and Cadastre.

BRUNEAU, D., SASSE, M. & MCCARTHY, J. (2002) The eyes never lie: The use of eye tracking data in HCI research.

CASTNER, H. & EASTMAN, R. (1984) Eye-Movement Parameters and Perceived Map ComplexityI. Cartography and Geographic Information Science, 11**,** 107-117.

CASTNER, H. W. & EASTMAN, J. R. (1985) Eye-Movement Parameters and Perceived Map Complexity-II. American Congress on Surveying and Mapping 0094-1689/85$2.50. The American Cartographer.

CHANG, K. T., LENZEN, T. & ANTES, J. (1985) The Effect of Experience on Reading Topographic Relief Information: Analyses of Performance and Eye Movements. Cartographic Journal, 22**,** 88-94

CHIN, C., LEE, S. & RAMEY, J. (2005) An Orientation to Eye Tracking in Usability Studies, Usability and Information Design.

ÇÖLTEKIN, A., HEIL, B., GARLANDINI, S. & FABRIKANT, S. (2009) Evaluating the effectiveness of interactive map interface designs: A case study integrating usability metrics with eye-movement analysis. Cartography and Geographic Information Science 36**,** 5-17.

COOKE, L. (2005) Eye tracking: How it works and how it relates to usability. Technical Communication, 52**,** 456-463.

CUDDIHY, E., GUAN, Z. & RAMEY, J. (2005) Protocol Considerations for Using Eye-Tracking in Website Usability

Testing. Vol. 2009.

DAVIES, C. & PEEBLES, D. (2007) Strategies for Orientation: The Role of 3D Landmark Salience and Map Alignment. Citeseer.

DELIKOSTIDIS, I. (2007) Methods and Techniques for Field-Based Usability Testing of Mobile Geo-Applications, MSc thesis. ITC, Enschede, the Netherlands, March 2007. Obtained on 8.1.2010 from http://www.itc.nl/library.

DELIKOSTIDIS, I. & VAN ELZAKKER, C. (2009) Geo-identification and pedestrian navigation with geo-mobile applications: how do users proceed? Location Based Services and TeleCartography II.

DESANTIS, R., ZHOU, Q. & RAMEY, J. (2005) A Comparison of Eye Tracking Tools in Usability Testing. Proceedings of the Society for Technical Communication Conference, Arlington, VA: STC

DITCHBURN, R. W. (1980) The function of small saccades, Vision Research. Vol. 20, 271-272, Pergamon Press Lid, Printed in Great Britain.

DUDA, R., HART, P. & STORK, D. G. (2001) Pattern Classification, New York, Wiley.

EHMKE, C. & WILSON, S. (2007) Identifying Web Usability Problems from Eye-Tracking Data. Proceedings of the 21st British CHI Group Annual Conference on HCI 2007: People and Computers XXI: HCI...but not as we know it. Vol.1. University of Lancaster, United Kingdom, The British Computer Society.

ELLIS, S., CANDREA, R., MISNER, J., CRAIG, C., LANKFORD, C. & HUTCHINSON, T. (1998) Windows to the soul? What eye movements tell us about software usability. Washington, DC: UPA Press.

GOLDBERG, J. (2003) WAM 2003. Eye tracking in usability evaluation: A practitioner's guide. The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research.

GOLDBERG, J. & KOTVAL, X. (1999) Computer interface evaluation using eye movements: Methods and constructs. International Journal of Industrial Ergonomics, 24**,** 631-645.

GOLDBERG, J., STIMSON, M., LEWENSTEIN, M., SCOTT, N. & WICHANSKY, A. (2002) Eye tracking in web search tasks: design implications. ACM New York, NY, USA.

GRAF, W. & KRUEGER, H. (1989) Ergonomic evaluation of user-interfaces by means of eye-movement data. Elsevier Science Inc. New York, NY, USA.

HANSEN, D. W. & HAMMOUD, R. I. (2006) An improved likelihood model for eye tracking. Computer Vision and Image Understanding 106 220-230.

HAYHOE, M. & BALLARD, D. (2005) Eye Movements in Natural Behavior. TRENDS in Cognitive Science, 9, pp. 188-194.

HENDERSON, J. (2003) Human gaze control during real-world scene perception. Trends in Cognitive Sciences, 7**,** 498-504.

HYONA, J., RADACH, R. & DEUBEL, H. (2003) The mind's eye: Cognitive and applied aspects of eye movements, London.

JACOB, R. (1995) Eye tracking in advanced interface design. Virtual environments and advanced interface design**,** 258-288.

JACOB, R. J. K. & KARN, K. S. (2003) Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. IN HYÖNÄ, J., RADACH, R. & DEUBEL, H. (Eds.) The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, Oxford, England, Elsevier.

JUST, M. A. & CARPENTER, P. A. (1980) A theory of reading: From eye fixations to comprehension. Psychological Review 87**,** 329-354.

KARN, K., ELLIS, S. & JULIANO, C. (1999) The hunt for usability: tracking eye movements. ACM.

KARSH, R. & BREITENBACH, F. W. (1983) Looking at looking: The amorphous fixation measure. In R. Groner, C. Menz, D. F. Fisher, & R. A. Monty (Eds.), Eye Movements and Psychological Functions: International Views (pp. 53-64). Hillsdale, NJ: Erlbaum.

LAMSWEERDE, A. V. (2000) Requirement Engineering in the year 00:A research Perspective. ACM computing surveys**,** 1-6.

LANKFORD, C. (2000) Gazetracker a software designed to facilitate eye movement analysis. The 2000 symposium on Eye tracking research & applications. New York, ACM.

MANHARTSBERGER, M. & ZELLHOFER, N. (2005) Eye tracking in usability research: What users really see. Citeseer.

MARSHALL, S. (2000) Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. Google Patents.

MAYR, E., KNIPFER, K. & WESSEL, D. (2009) In-Sights into Mobile Learning: An Exploration of Mobile Eye Tracking Methodology for Learning in Museums.

MENG, L. (2004) About Egocentric Geovisualisation.

MERCHANT, S. (2001) Eye movement research in aviation and commercially available eye trackers today. Eye Movement Summary-Assessing Human Visual Performance, Course at Department of Industrial Engineering, University of Iowa, USA.

NAJARI, R. (2009) Towards More Adaptive Pedestrian Navigation Systems. Enschede, the Netherlands, November 2009. Obtained on 8.1.2010 retrieved from http://www.itc.nl/library.

NAMAHN (2001) Using eye tracking for usability testing. Brussels, Namahn.

NEUROSCIENCE, A. (2002) A breadth-first survey of eye-tracking applications. Behavior Research Methods, Instruments, & Computers, 34**,** 455-470.

NEVALAINEN, S. & SAJANIEMI, J. (2004) Comparison of three eye tracking devices in psychology of programming research. Citeseer.

NGUYEN, H. L. (2010) Web visualization of trajectory data using open source web visualization library. Enschede, the Netherlands, Februrary 2010.

PERNICE, K. & NIELSEN, J. (2009) Eyetracking Methodology: How to Conduct and Evaluate Usability Studies Using Eyetracking.

POOLE, A. & BALL, L. (2005) Eye tracking in human-computer interaction and usability research. Encyclopedia of human computer interaction. Idea Group, Pennsylvania**,** 211-219.

POSNER, M. (1980) Orienting of attention. Month**,** 02.

PURVES, D., AUGUSTINE, H., FITZPATRICT, D., HALL, W., LAMANTIA, A., MCNAMARA, J. & WILLIAMS 3RD, S. (2004) Neuroscience 3rd edn (Sunderland, MA. Sinauer Associates.

RAYNER, K. (1998) Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124**,** 372-422.

REEDER, R., PIROLLI, P. & CARD, S. (2001) WebEyeMapper and WebLogger: tools for analyzing eye tracking data collected in web-use studies. ACM New York, NY, USA.

RICHARDSON, D. & SPIVEY, M. (2004) Eye-tracking: Characteristics and methods. Encyclopedia of Biomaterials and Biomedical Engineering. Marcel Dekker, Inc.

SALVUCCI, D. D. & GOLDBERG, J. H. (2000) Identifying fixations and saccades in eye-tracking protocols. Proceedings of the 2000 symposium on Eye tracking research \& applications. Florida, United States, ACM.

SHIC, F. & CHAWARSKA, K. (2008) The Incomplete Fixation Measure. ETRA. Savannah, Georgia, ACM 978-1-59593-982-1/08/0003 $5.00.

SORROWS, M. E. & HIRTLE, S. C. (1999) The nature of landmarks for real and electronic spaces. Spatial information Theory: Cognitive and Computational Foundations of Geographic Information Science. Springer-Verlag.

ŠPAKOV, O. (2008) iComponent-Device-Independent Platform for Analyzing Eye Movement Data and Developing Eye-Based Applications. Unpublished PhD Thesis, University of Tampere, Tampere, Finland.

STREEFKERK, J. W. (2006) Selecting, Combining and Tuning Evaluation Methods for Context-Aware Mobile User Interfaces for Professionals, TNO Defense, Security and Safety, Soesterberg, the Netherlands, Delft University of Technology, Delft, the netherlands.

TINKER, M. (1958) Recent studies of eye movements in reading. Psychological Bulletin, 55**,** 215-231.

TORSTLING, A. (2007) The Mean Gaze Path: Information Reduction and Non-Intrusive Attention Detection for Eye Tracking. Master’s Degree Project, Stockholm, Sweden.

TURANO, K., GERUSCHAT, D. & BAKER, F. (2003) Oculomotor strategies for the direction of gaze tested with a real-world activity. Vision Research, 43**,** 333-346.

VAN ELZAKKER, C., DELIKOSTIDIS, I. & VAN OOSTEROM, P. (2008) Field-Based Usability Evaluation Methodology for Mobile Geo-Applications. Cartographic Journal, The, 45**,** 139-149.

VAN ELZAKKER, C. & WEALANDS, K. (2007) Use and users of Multimedia Cartography. Cartwright, W. Peterson, MP, & Gartner, G. 2º Ed. Multimedia Cartography. ISBN-10**,** 3-540.

WEST, J., HAAKE, A., ROZANSKI, E. & KARN, K. (2006) eyePatterns: software for identifying patterns and similarities across fixation sequences. ACM.

WIDDEL, H. (1984) Operational problems in analysing eye movements. Theoretical and applied aspects of eye movement research, 21-29.
XIA, J., ARROWSMITH, C., JACKSON, M. & CARTWRIGHT, W. (2008) The wayfinding process relationships between decision-making and landmark utility. Tourism Management, 29, 445-457.
ZAMBARBIERI, D. (2005) Commercial uses of eyetracking, HCI 2005.

URLs

URL1 (2009) 'ASL Eye Tracking', http://asleyetracking.com/site.
URL2 (2009) 'Tobii Eye Tracking', http://www.tobii.com.
URL3 (2005), http://www.alexpoole.info/academic/lecturenotes.html.
URL4 (2001) 'Max Planck Institute for Psycholinguistics', http://www.mpi.nl/world/tg/eye-tracking/eye-tracking.html.
URL5 (2009) 'Usability Evaluation', http://www.usabilityhome.com.
URL6 (2009) 'Usability.gov', http://www.usability.gov.
URL7 (2009) 'Locarna Eye Tracking' , http://www.locarna.com
URL8 (2009) 'SMI Eye Tracking', http://www.smivision.com.
URL9 (2009) 'ISCAN Eye Tracking', http://www.iscaninc.com/.
URL10 (2009) 'Arrington Research, ViewPoint Eye Tracking', http://arringtonresearch.com/index.html.
URL11 (2009) 'SR Research Eye Tracking', http://sr-research.com/.
URL12 (2006) 'Chih-Hao Tsai's research page, Psycholinguistics, Psychology of Reading & Cognitive Science', http://research.chtsai.org/papers/scanpath-compression.html.
URL13 (2009) 'Interactive Minds, Eye Tracking Solutions', http://www.interactive-minds.com.
URL14 (2009) 'Eyetellect, LLC', http://www.gazetracker.com.

Appendices

Appendix A: Pseudo-codes for common ET clustering algorithms (Salvucci and Goldberg, 2000)

A. The VT algorithm:

1. Calculate point-to-point velocities for each point in the protocol.
2. Label each point below velocity threshold as a fixation point, otherwise as a saccade point.
3. Collapse consecutive fixation points into fixation groups, removing saccade points.
4. Map each fixation group to a fixation at the centroid of its points.
5. Return fixations.

B. The HMM algorithm:

1. Calculate point-to-point velocities for each point in the protocol.
2. Decode velocities with two-state HMM to identify points as fixation or saccade points.
3. Collapse consecutive fixation points into fixation groups, removing saccade points.
4. Map each fixation group to a fixation at the centroid of its points.
5. Return fixations.

C. The DT algorithm:

1. While there are still points.
2. Initialize window over first points to cover the duration threshold.
3. If dispersion of window points <= threshold.
4. Add additional points to the window until dispersion > threshold.
5. Note a fixation at the centroid of the window points.
6. Remove window points from points.
7. Else.
8. Remove first point from points.
9. Return fixations.

D. The MST algorithm:

1. Construct MST from protocol data points using Prim's algorithm.
2. Find the maximum branching depth for each MST point using a depth-first search.
3. Identify saccades as edges whose distances exceed predefined criteria.
4. Define the parametric properties (μ, σ) of local edges, identifying saccades when an edge length exceeds a defined ratio.
5. Identify fixations as clusters of points not separated by saccades.
6. Return fixations.

E. The AOI algorithm:

1. Label each point as a fixation point for the target area in which it lies, or as a saccade point if none.
2. Collapse consecutive fixation points for the same target into fixation groups, removing saccade points.
3. Remove fixation groups that do not span the minimum duration threshold.
4. Map each fixation group to a fixation at the centroid of its points.
5. Return fixations.

Appendix B: Pre-Selection Questionnaire for Potential Test Participants

Dear participant,

My name is Rozita Razeghi and I am a Geoinformatics MSc student who is currently completing my thesis under the title of 'usability of eye tracking as a user research technique in geo-information processing and dissemination'. This announcement is to request for your kind participation in a case study for my research. The followings include (A): a brief introduction to my research, and (B): some background information which I need in order to create your profile as one of the test participants.

(A) An introduction to the research:

This thesis investigates a rather new user research method called 'eye tracking' in the geo-domains. Eye tracking is done with a device which tracks people's eye movements to show where they are looking at. In this thesis a case study is performed which contributes to improving the usability of a pedestrian navigation system.

For this purpose, different methodologies can be applied during different phases of developing the system. One of the methods that can be applied in the first phase of design, is eye tracking. This case study applies and evaluates eye tracking and another new method for usability testing of a geo-mobile application to find the most usable method for development of this application. However, this case study does not aim at testing the mobile geo-application itself, but the combination of different methods and techniques in order to come up with the most informative usability testing methodology which could be later used for development or evaluation of the prototype of a mobile geo-application. In this survey, users are categorized into three different groups. Each group includes six participants who use the same method. Participants will be asked to do a navigation task using a PDA (geo-mobile application) through a predefined path e.g. going from 'A' to 'B'. The task will clearly be explained to the participants. The gathered user data will be analyzed later to determine the most informative method. Depending on the group that the participant is assigned to, different techniques will be used. The main techniques include eye tracking and video/audio recording which are combined with thinking aloud, questionnaire and interview depending on the participant's group.

All the surveys will take place in Enschede, not much far from ITC. Participants have to walk there from ITC and they will participate only once. The whole procedure will not take longer than 1.15 hours.

(B) Participant background information:

This questionnaire is aimed at selection of users such that they can form three homogeneous groups for comparison purposes. The personal information provided here and through the whole experiment process will be kept confidential. Each participant would be referred to via an ID.

I would like to thank you in advance for taking part in this questionnaire and I would like to know about your kind participation. In case you agree to participate in this research user experiment, please indicate the following information:

1.
   a) Name / Surname:
   b) Sex: □ M □ F
   c) Age:
   d) Nationality:
   e) Current education status:
   f) Do you wear contact lens: YES / NO

2. How long have you been in Enschede?

3. How much are you familiar with central part of Enschede? Please indicate your familiarity rate based on the template:
   □ Poor □ Modest □ Good
4. Do you have any experience with paper map?
   □ Yes □ No
5. If the answer to the previous question is positive, please rate your knowledge:
   □ Poor □ Modest □ Good
6. Do you have any experience with digital map?
   □ Yes □ No
7. If the answer to the previous question is positive, please rate your knowledge:
   □ Poor □ Modest □ Good
8. Do you have any experience with GPS systems?
   □ Yes □ No
9. If the answer to the previous question is positive, please rate your knowledge:
   □ Poor □ Modest □ Good
10. Do you have any experience with hand held devices like PDA or Smartphone?
    □ Yes □ No
11. If the answer to the previous question is positive, please rate your knowledge:
    □ Poor □ Modest □ Good
12. Do you have any experience with a mobile navigation application (like PDA or smart phon)?
    □ Yes □ No
13. If the answer to the previous question is positive, please rate your knowledge:
    □ Poor □ Modest □ Good
14. If the answer to question 12 is positive, please explain what application did you use? □
    □TomTom □ Route 66 □ iGo my way □ Destinator □ Others
15. If the answer to question 12 is positive, please indicate what you used it for?

16. If the answer to question 12 is positive, please mention the drawbacks you found in the application.

- Please leave your email and phone number, so that I can contact you in order to arrange a survey date/time comfortable for you.
  Participant phone number: Participant email address:

- If you have any question related to research, please contact me at:
  My cellphone: My email: razeghi21756@itc.nl

Appendix C: Post-Test Semi-Questionnaire

Please answer the following questions:

1. Describe your impression on using this field-based method (ET and TA technique) for the navigation and selection of salient LMs. Please rate your satisfaction with using the technique based on template.
   (1= very poor; 2= poor; 3= normal; 4= good; 5= very good)
   □1 □2 □3 □4 □5

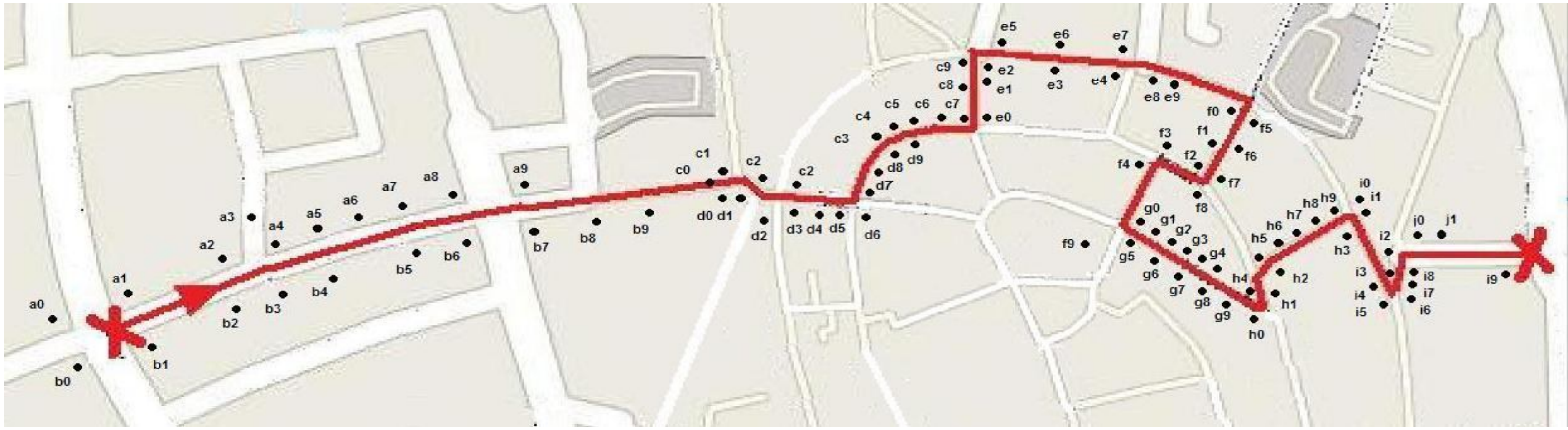
2. Did you experience any limitations or drawbacks in the test, due to using the ET technique?

3. Were you satisfied with using the ‘thinking aloud’ technique?
   □1 □2 □3 □4 □5

4. Did you experience any limitations or drawbacks in the test, due to using the ‘thinking aloud’ technique?

5. What suggestions do you have to improve the applied technique for obtaining user data information?

APPENDIX D: List of the LMs on the Pre-defined Path



C= Café
R= Restaurant
S=Shop

a0: Saxion University
a1: DTZ zadelhoff
a2: Notariskantool
a3: Stationplein
a4: JPR advocation
a5: Oosterhuis
a6: NAR akyla
a7: Emtes
a8: ATCL
a9: Tempo-team
b0: De groote schehuur
b1: Natuur meseum
b2: Vitvaartcentrum vredehot
b3: Staal bankiers
b4: Fortis bank
b5: SNS bank
b6: Parking place
b7: Snelderzijlstra
b8: Mevr baeken
b9: Uitzenbureau
c0: Under-ground bike repair
c1: Brammeter
c2: Bookshop
c3: C-Timeless
c4: C- Xotix
c5: C- Pjip
c6: Apres ski club
c7: C- Bieren
c8: C- Sate
c9: R- Detropen
d0: C- Chaplin
d1: R-Los Ponchos
d2: R-Turkish ..
d3: Bookshop
d4: S-Dynabyte
d5: S-Mobile action
d6: C-Talamini
d7: S-Head & smart
d8: C-Black sea
d9: C-Sahara
e0: R-Fred & douwe
e1: Cineast bioscoop
e2: ATM bank
e3: R-Humphrays
e4: C-Dance bar
e5: Muziekcentrum
e6: 2 old trees
e7: Muziekcentrum
e8: S-American gitar
e9: Bank
f0: C-The edge
f1: S- Chaplin
f2: R-Jour Mij Wijnhurs
f3: Church_ Revelation
f4: R- Buuvrow
f5: S- Capelli kappers
f6: Empty space
f7: S- Lampenpaleis
f8: R-?
f9: Church
g0: R- Pinocciou
g1: S- Van dor poel
g2: S- Faresjueweler
g3: S- Rosee
g4: S- M
g5: S- Gunnes
g6: S- Hip
g7: S- Tabito
g8: S- Merkkleding
g9: S- up=op
h0: S- China shop
h1: S- Laximi
h2: S- Koelink
h3: C- Mix
h4: S- Stolker
h5: S- ?
h6: S- Scooter
h7: S- Noppies
h8: S- Plato
h9: S- Avenue
i0: Jumbo-market
i2: S- China 1000
i3: S- Italiana
i4: S- Britain
i5: S- Het broekenhuis
i6: S- Emma
i7: S- Toko
i8: C-?
i9: Church
j0: S- Sea design
j1: C-?

Appendix E: Pre-Selection Questionnaire for Potential Test Participants

Dear Sir / Madam,
We are two Geoinformatics MSc students (Hoang Long Nguyen and Rozita Razeghi) who are currently completing our thesis under the titles of: 1. 'Web visualization of trajectory data using open source web visualization library' and 2. 'Usability of eye tracking as an user research technique in geo-information processing and dissemination'. This announcement is to request for your kind participation in a common case study required for our researches. The followings include (A): a brief introduction to our researches, and (B): some background information which we need in order to create your profile as one of the potential test participants.

(A) An introduction to the research:

1. The main objective of the first MSc research is to propose a framework for visual exploration of trajectory data on the Web using existing visualization libraries. Iceberg data are used as a case study. The result of this thesis is a Web application for visual exploration of iceberg data. This usability testing aims to investigate the usability of a prototype design solution for visual exploration of iceberg data.

2. The second thesis investigates a rather new user research method called 'eye tracking' in the geo-domains. Eye tracking is done with a device which tracks people's eye movements to show where they are looking at. In this thesis, a case study is performed which contributes to improving the usability of a website designed for the visualization of the icebergs' trajectory data. For this purpose, different methodologies can be applied and one of the methods that can be applied is eye tracking. This case study applies and evaluates eye tracking alone, and also eye tracking combined with audio-video recording, in an usability testing to find the most usable method for the improvement of this application. However, this case study does not aim at evaluating the geo-application itself, but the combination of different methods and techniques in order to come up with the most informative usability testing methodology.

In our joint case study users are categorized into two different groups. Each group includes 4 participants who use the same research method. Participants will be asked to do some tasks using the website. The tasks will be clearly explained to the participants. The gathered user data will be analyzed later to determine the most informative method. Depending on the group that the participant is assigned to, different techniques will be used. The applied techniques include eye tracking, thinking aloud, questionnaire and interview.

The test will take place in Enschede, at the Campus of Twente University. Participants will go there from ITC and come back by bus, and they will participate only once. The whole procedure including the transportation will take around 1:30 hours. The user research will take place on Monday, 1st of February at a time that suits you most. You will help us with your participation, but we also think this is a nice opportunity for you to become acquainted with modern ways of doing user research.

(B) Participant background information:

This questionnaire is aimed at selection of users such that they can form two homogeneous groups for comparison purposes. The personal information provided here and throughout the whole experiment process will be kept confidential. Each participant would be referred to via an ID.

We would like to thank you in advance for taking part in this questionnaire and we would like to know about your kind participation.

Are you willing to participate in this user research? YES / NO

1. Personal information

a) Name / Surname:
b) Sex: □ F □ M
c) Age:
d) Nationality:
e) Current education status:
f) Do you wear contact lens: YES / NO
g) Do you wear bi-focal glasses: YES / NO

2. How familiar are you with trajectory data or movement data? Please indicate your familiarity rate based on the template:
   (1 = very poor; 2 = poor; 3 = normal; 4 = good; 5 = very good)
   □1 □2 □3 □4 □5
3. How much do you know about icebergs and their characteristics?
   □1 □2 □3 □4 □5
4. How familiar are you with using an interactive map?
   □1 □2 □3 □4 □5
5. How familiar are you with using animation?
   □1 □2 □3 □4 □5
6. How familiar are you with using websites application?
   □1 □2 □3 □4 □5

- Please leave your email and phone number, so that we can contact you in order to arrange a date/time comfortable for you.
  Participant phone number:
  Participant email address:

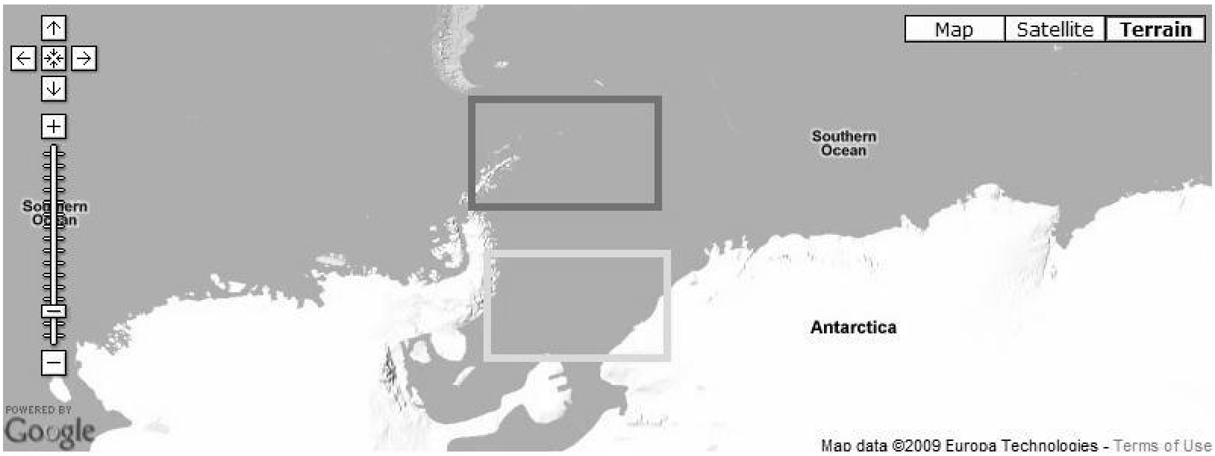
- If you have any question related to research, please contact us at:
  Our emails: razeghi21756@itc.nl, nguyen17801@itc.nl

Appendix F: Test Questions

The evaluation is based on the use of functions and views while the participants try to answer questions. The questions are separated into 3 groups according to the Visual Information Seeking Mantra 'overview, zoom and filter, details on demand':

Group 1: Questions related to overview

1. In which zone there were more icebergs appearances, zone 1 (inside the blue box) or zone 2 (inside the yellow box) (see figure below)?



2. In which year were more appearances of icebergs?
3. Which iceberg has the longest lifetime?

Group 2: Filtering and zooming

1. List 3 icebergs that had slow average speed (less than 10 km/per day)?
2. Where is the iceberg A01 at 11/01/1979? (provide the latitude and longitude position)
3. Compare the sizes of icebergs A20A and A20B at their appearances. Is A20A bigger than A20B?

Group 3: Details on demand

1. Where and when did the iceberg A15 disappear? (Provide the latitude and longitude position)
2. How much did the size of iceberg A19 decrease between its appearance and its disappearance?
3. Did the iceberg A20A's movement cross the A20B's movement?

Appendix G: Post-Test Semi-Questionnaire

Please answer the following questions:

1. Were you satisfied with using the ET technique? Please rate your satisfaction with using the technique based on template.
(1= very poor; 2= poor; 3= normal; 4= good; 5= very good)
□1 □2 □3 □4 □5

2. Did you experience any limitations or drawbacks in the test, due to using the ET technique?

3. Were you satisfied with using the ‘thinking aloud’ technique?
□1 □2 □3 □4 □5

4. Did you experience any limitations or drawbacks in the test, due to using the ‘thinking aloud’ technique?

5. Do you have any suggestions for the improvement of this user research method?