18

Simple Guidelines for Testing VR Applications

Livatino Salvatore[1] and Koeffel Christina[2]
[1]*Electronic, Communication and El. Engineering, University of Hertfordshire*
[2]*Center for Usability Research and Engineering*
[1]*United Kingdom,* [2]*Austria*

1. Abstract

In recent years the number of virtual reality (VR) applications and devices employed in companies as well as in research facilities has increased remarkably. The applications developed call for human-computer interaction, which in turn calls for system and usability evaluations. These evaluations are usually conducted by means of measurement of human behaviour including aspects of perception, action, and task-performance. Traditionally, evaluations are held in specially designed usability laboratories but the new tendency requires researchers and even students that are non-experts in the field of usability, to evaluate applications. Hence, these studies take place at the premises of laboratories at universities or research institutes. This raises the question of whether non-experts are able to conduct evaluations in a professional manner. Furthermore, the evaluation issue calls for multi- and inter- disciplinary collaboration, where technical expertise is combined with humanistic knowledge and methodology. Several experts in the field of VR, as well as in the field of usability studies, call for producing helpful guidelines in order to be able to evaluate VR applications. This chapter gives an overview of the problem and introduces a guideline for evaluations of VR applications which aims to assist researchers in evaluating VR devices and installations,and in particular usability novices. This chapter also aims to facilitate multidisciplinary activities through the use of an evaluation guideline which would be simple and focused on VR. The applicability and usefulness of the proposed guideline are based on the authors' experience in supervising and coordinating several university student projects related to the development of VR visualization technologies and applications. Furthermore, the guideline has been tested through several case studies where students were asked to evaluate their final year projects. The results showed how the proposed guideline could successfully be employed by non-experts to carry out pilot and formal evaluations. Therefore, this chapter is expected to represent a valuable reference for students and non-usability experts who design VR systems or applications and wish to run a user study to assess usefulness and general usability characteristics of their products.

2. Introduction

In recent years VR applications started to be used more commonly. If a few years ago only big institutions had the budget to acquire devices such as a CAVE or a head mounted display, today the spread of interactive applications and their commercial use has opened a wide market accessible to a larger range of activities and budgets. This has also raised interest in the ways of using VR devices. Research institutions are exploring new possibilities and new interaction tools, to support and enhance the ideal use of VR facilities.

There is a large range of possible applications in VR, not only for product development and laboratory simulation, such as computer-aided design, data visualization, training, but also for entertainment purposes, such as computer games, art and tourism. The full potential of those techniques has not entirely been explored by now, e.g. only few games suitable for VR devices are available so far.

The VR evolution also raises the question of usability to assess the relevance and user friendliness of the proposed applications. Especially since each VR device has a diverse operational area, there is a large range of possible uses depending on the given circumstances. In order to assess the overall usability of the system and to figure out which kind of device and system setting should better be used on what application, usability evaluations of systems are conducted.

In the last 20 years, since the introduction of usability to a wider audience (also comprising non experts), amongst others by Jakob Nielsen (Nielsen, 1993), the development of usable applications whose usability is evaluated through user-tests has increased in importance. Companies offering consultancy in this area have been established, e.g. USECON[1], which most of the time dispose of professional usability laboratories. Nevertheless, only few research establishments, especially from an interdisciplinary background, have access to such facilities.

The number of people developing VR devices increases steadily. Amongst them there are computer-scientists and engineers, including also a remarkable number of students, who are often required to conduct user studies by themselves. Most of those technically educated people are not specialists in conducting evaluations.

Unfortunately, at the current state little work can be found in form of helping instructions or guidelines for the specific area of VR. Although evaluations and usability in general are a common and well known area and there are several books available on this topic, it has to be considered that VR is a modern and not standardized subject. Furthermore, students usually do not have a large amount of time for developing their applications or doing their research. Big companies or universities specializing in this field have more resources at hand. Students typically have up to one year of time to develop and test a project. Therefore most of the guidelines introduced have to be simplified in order to be able to be applied in a reasonable amount of time. Nevertheless, they have to be as accurate as possible.

The purpose of this chapter is to describe a simple guideline for running usability studies with a focus on VR applications. The guideline consists of a number of directives exposed with a certain level of detail, while still being a compact set (as required for a handbook), and of easy understanding.

This chapter is expected to represent a precious reference for students and non usability experts that undertake the design of a new VR application and wish to run a pilot or formal user study to assess effectiveness and general usability characteristics of their product.

Section 3 provides an overview of related work which supports the argumentation for a simplified and focused guideline. Section 4 introduces the guideline concept in terms of a multidisciplinary facilitator. Section 5 presents the objective and approach of the chapter.

[1] USECON, The Usability Consultants, http://www.usecon.com/

Section 6 presents the suggested basic set of evaluation guidelines to be considered when performing a user study. A case study where the proposed guideline has been adopted is described in section 7.

3. Related Work

Renowned specialists in the field of VR, who are conducting usability studies and perform research in this field, call for a general guideline for evaluations in VR.

Joseph Gabbard states in (Gabbard, 1997) the urge for a handbook or more professional help for conducting usability studies in the field of VR. In his thesis, Gabbard proposes a taxonomy for conducting user studies connected to VR. Gabbard offers a very precise and detailed list of inputs for conducting usability studies. He introduces an approach which starts at the beginning of the product life-cycle and comprises assets for the tasks the participants should conduct as well as for the specific applications that are employed for the usability study. Although written in 1997, most of the inputs are still valid, nevertheless, there is no section about the state-of-the-art included in his thesis - a point that he tried to cover in his following work.

In particular, Gabbard et al. published a paper in 1999, describing a new methodology treating the usability process of products employed in the field of VR (Gabbard et al., 1999). They divide the interaction development into behavioural and constructional development. Because of this user-centered and developer-centered approach, they describe a four-step usability process including user task analysis, expert guideline-based evaluation, formative user-centered evaluation and summative comparative evaluations. Gabbard et al. further mention the lack of professional guidelines in VR. Although they describe a well-structured process, the guidelines still consist of expert-based approaches that are hard to be executed by non-experts in the field of usability.

Bach and Scaping, (Bach & Scaping, 2004) are also very involved in this area and point out the lack of a guideline that not only treats one VR device, but is a general handbook for all VR facilities available. Further they mention that usability tests in the field of VR could get very complicated, since the technology used is very complex. In general, they give a good overview of the main obstacles one might encounter while conducting usability studies in this area. Bach and Scaping specify the most significant differences between common user studies and user studies accomplished in the field of VR. They also point out that most of the VR devices are multi-user compatible, whereas regular 2D devices are not. Since the ergonomic knowledge in the field of VR is very rare, they suggest combining some already existing methods of evaluations with new ideas.

In his paper about usability of VR systems, Timothy Marsh criticizes the absence of a general guideline for 3D applications (Marsh, 1999). He proposes further research to investigate the precise definition of VR, to detect the differences in the evaluation methods from normal 2D usability studies, to investigate existing research approaches and to develop studies for a future research plan.

Sutcliffe and Gault (Sutcliffe & Gault, 2004) introduce a new possibility of heuristic evaluations for VR applications. Their method is based on the heuristics introduced by Nielsen (Nielsen, 1993). They assembled a set of 12 heuristics and conducted 2 test studies using them. The results show, that not all heuristics are suitable for all kinds of applications and devices. Sutcliffe et al. consider the heuristics developed to be a new and important

extension to expert-based human-computer interaction (HCI) evaluation methods and they deem it to be superior to Nielsen's original approach concerning VR applications.
In 2006 Karaseitanidis et al., (Karaseitanidis et al., 2006), used the VIEW-IT tool (view inspection tool) developed by Tromp and Nichols (Tromp & Nichols, 2003) for a new kind of evaluation method, following the "VIEW of the Future" project. They decided not only to rely on traditional questionnaires, but to use psycho-physiological and neurophysiologic measures like the estimation of mental workload, stress, strain, level of cognitive performance, alertness and arousal. For assessing these factors, which are not able to be measured with "common" usability methods, they made use of electroencephalography (EEG), event-related potentials (ERP), electrocardiography (ECG), blood pressure and more. Furthermore they made a socio-economic assessment to identify future domains for VR facilities.
As can be seen from the above presented works, the methods and approaches introduced almost exclusively focus on expert-based evaluation methods and are therefore not suited for novices in the area of usability.

4. A Multi-Disciplinary Subject

Measurement of human behaviour is a main subject of the humanistic community (e.g. psycho-physiologists, psychologists), whose studies have recently also targeted VR related topics (e.g. presence, immersion). The humanistic community typically lacks technical knowledge which sometime leads to proposing inapplicable studies or studies of difficult comprehension by many among the technical people.
VR technology is, on the other hand, mostly developed by the technical community (e.g. computer scientists, engineers). However, in this community there are many not familiar with conducting human-based evaluations and therefore struggling with designing a systematic usability study, hence, desiring simple guidelines or, in other words, an evaluation "handbook".
While multi- and inter- disciplinary research activities are being promoted and carried out both at research and educational level, a synergic interaction between the technical and humanistic communities is still out of reach for many researchers who have typically been educated monodisciplinary.
Interaction between different communities has often led to contrasts, misunderstandings and slow development. The authors of this chapter and their department colleagues have experience with problematic situations due to multidisciplinary cooperation in recently established interdisciplinary educations, such as Medialogy[2], and in projects at European level, e.g. Benogo[3], Puppet[4].
The aim for this guideline is also to support multidisciplinary activities by functioning as facilitator of multidisciplinary interaction when performing evaluation studies in VR. In

[2] Medialogy is a new interdisciplinary study at Aalborg University, Denmark. http://www.media.aau.dk.
[3] Benogo (Being There Without Going), EU-FP6 RTD (IST-FET) project. Coordinator: Aalborg University, Denmark, 2002-2005. http://www.benogo.dk
[4] Puppet (The Educational Puppet Theatre Of Virtual Worlds), EU-ESPRIT-LTR project under i3 – (Experimental School Environment). Coordinator: Aalborg University, 1998-2001. http://www.cvmt.dk/projects/puppet.

particular, by reducing interaction among different disciplines when this may not be needed. For example, when non-usability experts aim at performing pilot/formal performance evaluations or when evaluations are planned after product development and not earlier in the process (such as for industrial products).
Although user studies in VR should generally involve multidisciplinary collaborations because of their complexity, the authors of this chapter believe that it is beneficial to provide VR technical developers with a handbook containing a precise guideline on how to perform usability tests. A concept also supported by the experts mentioned in the previous section.

5. Aim and Approach

The aim of this chapter is a simple evaluation guideline specifically targeting usability studies in VR. The guideline is designed for "non-experts" to provide them with a simple but scientifically valid reference on how to run and design usability evaluations. We propose the guideline to be achieved from the analysis of representative and well assessed scientific work in the field, e.g. articles from major international journals and conferences.
We have researched and analyzed the state of the art in evaluations connected to VR in some of our
previous work, (Koeffel, 2007; Koeffel, 2008). In particular, more than 30 representative literature works concerned with evaluation of VR displays have been analyzed, and 15 of them have been chosen for a deeper study (Bayyari & Tudoreanu, 2006; Christou et al., 2006; Cliburn & Krantz, 2008; Demiralp et al., 2006; Elmqvist & Tudoreanu, 2006; Fink et al., 2004; McMahan et al, 2006; Ni et al., 2006; Patel et al., 2006; Qi et al., 2006; Sutcliffe et al., 2006; Takatalo et al., 2008; Vinayagamoorthy et al., 2006; Wang et al., 2006; Wilcox et al., 2006). The analyzed papers represent a determined cross-section of the mostly treated areas in VR and are related to the analysis of different visual display technologies in VR applications.
Since the field of VR represents a wide area which is surely hard to cover with a general guideline, we decided to leave out acoustic and haptic devices. Furthermore, not all visual displays are considered, e.g. head mounted displays, since it was our goal to show case studies through experimentation with available facilities (though, including a large panoramic wall and a 6-sided CAVE).
Generally, one major issue in visual display technology in VR, covered in some of the considered papers, is the display size. Among other topics of interest: exploration and remote driving, information rich VRs, influence of the real world and occlusions in VR, and use of different VR facilities.
The Fishtank VR facility is evaluated against the CAVE in two different types of user studies conducted by Demiralp et al., (Demiralp et al., 2006). Ni et al. investigate the connection between display size and task performance in information-rich VR, (Ni et al., 2006). Bayyari and Tudoreanu investigate the impact of immersive VR displays on user performance and comprehension (Bayyari & Tudoreanu, 2006). Cliburn and Krantz conduct a user study in order to assess the impact of stereoscopic visualization and multiple displays on user performance (Cliburn & Krantz, 2008).
The users' performance in real and virtual environments is investigated by Fink et al. (Fink et al., 2004). A similar study is conducted by Sutcliffe et al., who compare interaction in real world and virtual environments (Sutcliffe et al., 2006). In (Wang et al., 2006), Wang et al. evaluate the effects of real world distraction on user performance in VR. The effectiveness of occlusion reduction techniques is investigated by Elmqvist and Tudoreanu in (Elmqvist &

Tudoreanu, 2006). Using the results of an empirical study, Qi et al. develop guidelines to ease the choice of display environment for specific volume visualization problems (Qi et al., 2006).

The performance and usability of different 3D interaction devices is investigated by Patel et al. (Patel et al., 2006). The importance of posture and facial expressions of virtual characters in virtual environments is explored by Vinayagamoorthy et al. (Vinayagamoorthy et al., 2006). Wilcox et al. evaluate whether or not people perceive a violation of their interpersonal space when using VR (Wilcox et al., 2006). Takatalo et al. present a framework for measuring human experience in virtual environments in (Takatalo et al., 2008). McMahan et al. introduce a study that separates 3D interaction techniques and the level of immersion since it is very difficult to compare different VR systems, (McMahan et al, 2006). The development and evaluation of a large-scale multimodal VR simulation suitable for the visualization of cultural heritage sites and architectural planning is described by Christou et al. (Christou et al., 2006).

The most significant information about the user studies conducted in the analyzed papers had been collected, classified, and used as a base for our analysis.

The data collected comprehends:

- The goal of the user study.
- The number of participants.
- Data connected to the participants such as their vision or attitudes.
- The setup in use.
- The number of tasks to complete, their nature and the completion time.
- The statistical and graphical types of evaluation.

All the collected data has then been evaluated statistically.

The idea of a "handbook" together with its content have then been developed, (Koeffel, 2008), presented to the research community (Livatino & Koeffel, 2007), and further improved based on the acquired experience.

The proposed guideline merges existing traditional approaches in evaluation and the state of the art in evaluating VR applications. For the traditional part especially Nielsen (Nielsen, 1993), Rubin (Rubin, 1994) and Faulkner (Faulkner, 2000) appeared to be relevant.

The guideline addresses two areas:

1. General suggestions on how to design usability studies.
2. VR specific aspects of usability studies.

The general recommendations of the first part can be understood as the basics of user testing, which can also be found in the literature, such as (Faulkner, 2000; Nielsen, 1993; Nielsen & Mack, 1994; Rubin, 1994; Sarodnick & Brau, 2006). The proposed guideline comprises a summary of the suggestions offered in the above mentioned sources and because of the generality of this part, the provided information can also be adapted and applied to different areas of VR.

The second part includes VR specific aspects such as the number of participants employed in user studies in the field of VR, the kind of questionnaires used or statistical measures employed to obtain results, etc.

The Guideline has been tested through user studies, (one of them is briefly described in section 7). The results and experience obtained from those studies combined with the authors' experience in conducting user tests, have allowed for improving and completing the guideline to the stage presented here.

The proposed guideline contains the basis of a recommendation on the most important steps for designing, planning and conducting a usability study in the field of VR. It is accompanied by a case study to facilitate implementation and conceptual understanding.

6. Evaluation Guidelines

This section describes the proposed set of directives for the evaluation of VR applications. The following paragraphs are divided into sub-sections addressing specific aspects. The content of the description is based on selected literature in the field of VR that we have investigated. The test designer is left with some freedom of choice depending on: the guideline specific aspects, application context, available time, and pre-determined objectives. To support the designer's decision in making a choice, the guideline often directly refers to the results of our investigation in specific aspects in terms of percentage of literature works.

6.1 Research Question

Before starting to build a setup for an evaluation, the research question for the usability study needs to be formulated. A general research question defining the purpose of the entire project should already exist; nevertheless a specific research question should be formulated for the special purpose of the evaluation. This defines the main subject of the study.
It is very important to create a strong and valid research question that summarizes the goal of the evaluation in only one sentence/paragraph.
It is essential that the purpose of the entire project as well as the evaluation is clear to everybody on the project/evaluation team. Additionally, the research question should help to formulate the hypothesis we want the project to be tested against.

6.2 Ethics

Since user tests are conducted with humans, it is essential to assure that there will be no harm to the participants and that their personal rights are maintained, (Burdea & Coiffet, 2003). Users' mental and physical health must not be at risk and they need to be informed about potential hazards. Furthermore, users have to be able to stop whenever they feel uncomfortable and desire the test to end.
Certain universities dispose of an ethical department that administrates all studies and evaluations conducted involving humans. In this case, the researchers have to apply to this committee and do have to obey certain rules. If the university where the evaluation is supposed to take place does not dispose of such a department, ethical considerations have to be taken into account as well. Especially when there is no board reviewing the studies, one has to make sure that all ethical concerns are respected. Furthermore also legal considerations of the country where the study is planned,should be reviewed.
Virtual reality applications offer many possible risks to the participants of a user study, e.g. in cases when new devices are invented and tested or when existing devices have not entirely been tested for health risks. Additional hazards can appear through the use of HMD's, backpacks or laser diodes. Different mechanical devices in use, such as haptic tools can endanger the participants' health when applied incorrectly. Side-effects such as the occurrence of cybersickness need attention when using VR applications. They might even require a participant to stop the test.

6.3 Evaluation Method

At the very beginning of each user study it is important to choose and define the appropriate evaluation methods applicable to the setup to be tested. According to Jakob Nielsen these are: performance measures, thinking aloud, questionnaires, interviews, logging actual use and user feedback. These evaluation methods can also be applied in a combined version.

Depending on the time when the evaluation takes place and the kind of data collected, one can distinguish between formative and summative user studies. Formative usability evaluations usually take place several times during the development cycle of a product to collect data of prototypes. Typically summative evaluations are applied at the end of a project, for example, to compare different products. Formative user studies are rare in VR. According to our statistical evaluation, all of the 19 user studies investigated were conducted as summative studies.

When comparing two or more different VR devices/applications (summative evaluation), one can decide whether to use a within or between subjects design. Between subjects studies are more common in VR. A statistical analysis conducted in (Koeffel, 2008) has shown that a total of 61% of user studies in VR were designed as between subjects studies.

6.4 Setup

In our recommendations the setup is distinguished into the testing environment and the technological setup.

- **Testing environment**
  Evaluations conducted by students and academic researchers usually take place in the facilities of universities. In some cases universities dispose of their own usability labs for conducting evaluations, but in most of the cases the evaluations occur in computer labs or classrooms. Since classrooms are not always comfortable (and hard to find relaxing), while it is required that the participants feel at ease, it is very important to create a comfortable environment. It has to be avoided the presence of people that are not involved in the project, the presence of those running around hectically and preparing the evaluation, and other distractions such as loud noises.
  It is generally important not to give the participants unnecessary information about the project or to bias the results by telling the users some weaknesses or previous results. If the user study requires a test monitor logging data while the participants perform the testing, it is fundamental that he/she respects the participants' privacy by not sitting too close to them. Furthermore, any kind of stress and emotional pressure has to be kept away from the participants in order not to influence the results.
- **Technological setup**
  Student and research evaluations often base on an already finished project (summative evaluations), referring to the hardware and/or the software. Therefore the technological setup might already be given. Considering the different VR setups, it is very important to assure that all needed devices are at the test monitors' disposal on the day(s) of the usability study. Furthermore it is very important to test if the application, the software, and the data logging, are well functioning. Since VR devices and applications are still considered to be "new technology" they are sometimes unstable and tend not to work all the time. Hence, it is crucial to organize and reserve technical facilities and rooms, and to inspect the functionalities of the project to be tested.

6.5 Participants

Several fundamental elements of evaluations are related to participants. Before recruiting volunteers, it is very important to investigate the target population of the user study. Therefore users with the desired attributes such as age, gender, education, experience with VR, computer experience, gaming experience, visual abilities, etc. can be selected. Generally, it is advisable to test user groups with a great internal variance. Especially in the field of VR, it is utterly important to recruit users from different age groups, gender and experience.

The results of our research indicate that around 71 percent of the participants were male and only 29 percent were female, which means that men are participating in user studies in the field of VR twice as often as women. This could be acceptable in case of men being the main users of a VR product. In general, a careful selection of participants according to expected system users is important.

Concerning the number of participants, it mainly depends on the kind of user study conducted (i.e. formative or summative evaluation, between or within subjects design, etc.). Generally, usability experts (Faulkner, 2000; Nielsen, 1993; Nielsen & Mack, 1994; Rubin, 1994) hold that 2 to 4 participants suffice for conducting a representative pilot study, and 5 to 20 participants suffice for conducting a formal user study. Nevertheless, more recent approaches on evaluations in the field of VR have suggested testing a higher number of participants in order to obtain meaningful results.

A number of approximately 23 participants is suggested for within subject designs and 32 for between subject designs. In case of pilot studies, a minimum number of 6 participants is proposed. These figures are based on our literature analysis.

Participants are typically volunteers and/or they do not receive any financial compensation. Nevertheless, it is highly recommended to hand the participants a small token of appreciation after finishing the user study.

6.6 Forms

Forms are different documents that are handed to the participants during the course of a user study. Concerning the forms given to the participants, this guideline conforms to the traditional approaches introduced in (Nielsen, 1993; Rubin, 1994) and the results of the statistical analysis of relevant literature in (Koeffel, 2008). Therefore we recommend the use of the following forms:

- **Information sheet:** The information sheet (also called test script) provides an overview of the entire testing process. This form should be handed to the participants at the very beginning before the actual testing, and it should contain information about: the title of the project, names and contact information, introduction to the project, duration of the study, tasks to be completed, and the possibility to withdraw from the study at any time. In 5 out of the 18 studies investigated, the participants have reported to have received written or displayed information before the testing process.
- **Consent form:** The consent form states that the researchers are allowed to use and publish the data collected during the user study. This may also include eventually taken pictures or videos. It is a reassurance for the participants that their data will not be used for any other purpose than the one explained in the consent form and/or in the information sheet. For the researcher this form is a legal reassurance that he/she is allowed to use and publish the obtained data.

- **Questionnaires:** Generally questionnaires should contain the information required by the research question which is not possible to be collected automatically through data logging and performance measures. Therefore, mostly subjective qualitative data is collected using questionnaires. Special issues should be treated in questionnaires in order to emphasize the conclusion and the results of the data collection. Questionnaires can provide answers about personal feelings or preferences. We distinguish among: screening, pre-test, device, post-test, background, presence, simulator sickness, and the EVE-experience questionnaire.
- **Task scenarios:** It might be necessary to provide participants with a task scenario (describing each step in detail) for each task he/she should complete. This allows every participant to gain the same amount of information. Furthermore it clarifies the knowledge necessary to complete a given task.
- **Data collection forms:** Experience has shown that it is not always sufficient to auto-log data using software. Sometimes it is necessary that the test monitor writes down notes or information during a task session. This can be additional information such as time or estimates expressed by participants.
- **Thank you form:** In addition to the possible personal gratification that participants may receive by taking part in a user study, a thank you letter should also be handed to them. This is important in order to formally thank the participants and tell them where to find further information about the progress of the project, e.g. published papers.

The forms should be adapted to the needs of the user study. Generally we suggest the employment of semantic differentials as answering options.

6.7 Schedule

It is essential to estimate the overall completion time per participant and to prepare a schedule showing the sequence of participants and their assigned tasks. In particular, the schedule should include: timing of the single tasks, overall completion time, the sequence of the tasks per participant, possible breaks, time needed for introduction and debriefing, room locations, etc.

The studies analyzed indicate an overall completion time per participant that ranges from 23 to 240 minutes with an average completion time of 45 minutes. This time includes the time from when a participant arrived at the testing facility until the time he/she left.

In the field of VR it is very important to keep the single task sessions as well as the overall completion time as short as possible. A maximum of 30 minutes per task is recommended by Bowman et al. (Bowman et al., 2002). Too long sessions might cause exhaustion of the participants and side effects such as cyber-sickness, which could negatively affect the results. It is important to counterbalance the sequence of the single tasks in order to avoid learning effects and biasing of the results.

6.8 Test Monitor and Spectators

The role of each person present during the user study has to be predefined. Especially the test monitor should be well instructed and capable to serve his/her purpose.

The test monitor is present during all parts of the usability study and interacts with the participants. If possible somebody who has ground knowledge in usability (especially evaluations) should be employed as test monitor. In case that there is no expert in usability available, the person in the role of test monitor should acquire basic knowledge in this area.

The test monitor should be able to comfortably interact with the participants, which requires an open and friendly personality (i.e. a "people-person"). It is also important that the test monitor does not get too close to the participants physically as well as mentally, to give them some privacy.

In case other people than the test monitor and the participant, are present during a test session, e.g. technical staff, VR project development team, spectators, etc., they should be introduced to participants at the beginning and the roles of the spectators need to be defined clearly. Generally, the number of spectators during a testing session should be kept small since they tend to make the users nervous. If not part of the research question, spectators should avoid talking during task sessions. This is especially crucial for VR applications, since distractions such as loud noises might disturb the sense of presence.

Since VR systems are still considered new technology and unstable, it might happen that the participant gets frustrated because something is not working properly or it is very difficult to accomplish. In such a case, the test monitor should not judge the participant or the system by expressing that e.g. "this error always occurs" or otherwise by negatively influencing the user. The test monitor should encourage the participant to go on as long as possible.

6.9 Test Plan

When forming the idea of conducting an evaluation, a test plan should be created. This document contains in principle every kind of knowledge necessary for the user study. The test plan describes the main content of the usability study. It serves as the basic document for communication to other people that might be involved in the user study (e.g. second test monitor).

Using the test plan every involved person knows the main principles and ideas behind the evaluation. Therefore open questions and misunderstandings can be clarified. Furthermore, the test plan describes the resources needed and gives an overview of the milestones already accomplished. A properly formulated test plan for user studies in the field of VR should contain the following items:

- **Purpose**: The purpose describes the research question and the main problems treated as well as the current state of the art of the project.
- **Problem statement/test objectives**: The problem statement treats the main issues and questions connected to the evaluation, e.g. the questions derived from the hypothesis.
- **User profile**: The user profile describes the target group and the participants to be acquired for the study.
- **Test design**: The test design includes decisions about the entire session of the usability study, such as the evaluation method, e.g. if doing a between or within subjects evaluation. Furthermore the test design specifically describes each single step during the user study, starting from the arrival of the participants until the time they leave.
- **Task list**: The task list describes every task and subtask that the participants will be asked to accomplish and on which VR device tasks are accomplished.
- **Test environment/equipment**: This section elaborates the test environment and equipment used in the test, e.g. VR devices and rooms needed.
- **Test monitor role**: The description of the test monitor role includes information about the test monitor and possible spectators.
- **Evaluation measures**: The evaluation measures should be described on a list enumerating all data collected during the user study (data logging, questionnaires, etc.).

- **Report contents and presentation**: This section gives a short preview on the data contained in the final test report and the presentation of the results obtained during the user study.

6.10 Pilot Study

It is generally recommended to perform a pilot study before testing a project in a formal user study. The pilot study should be conducted in the same way as the formal study and each participant should be treated as if he/she were in the formal study (including the forms to be used).
The pilot study is useful for removing errors from the project/setup, debug the test design, debug the experimental design, detect biased questions in the questionnaires, refine the questionnaires and detect the overall time necessary per participant. Furthermore, rooms and technical facilities should be tested of their functionality.
A minimum number of 6 participants is suggested, (see sub-section 6.5 for more details). In general, the more participants are tested, the more indicative the results are.
The pilot study is essential in case of problems that may not be predicted and only occur during the study.

6.11 Formal Study

In an ideal case, a pilot study has been conducted before the formal study and the results of the pilot study have been taken into account when planning and conducting the formal study. If required, additional tests could be conducted at the very beginning of the study in order to categorize the participants. Furthermore, a practice session should be administrated for all testing activities which need a test-user to become acquainted with system commands and behaviour. In our literature an average of 4.1 tasks are accomplished per participant in practice sessions.
In order to avoid negative side effects (such as motion sickness) and fatigue, long enough breaks should be held between the single task sessions.
Concerning the variables chosen for the evaluation, the most frequently used independent variables (factors) are related to: display type and size, chosen virtual environment, stereoscopic visualization, etc. The most popular dependent variables are subjective measures (obtained from questionnaires). The most used are completion time and response accuracy.
The number of possible tasks which users are asked to complete during an evaluation may vary largely. An average of 28.7 trials accomplished by one participant during a user study was detected in the analyzed works. The most frequently employed categories of tasks in our literature research may represent a typical example. These are:

1. Navigation (Bayyari & Tudoreanu, 2006; Cliburn & Krantz, 2008 ; Fink et al., 2007; Ni et al., 2006; Patel et al., 2006; Takatalo et al, 2008, Vinayagamoorthy et al., 2006; Wang et al., 2006).
2. Manipulation (Christou et al., 2006; McMahan et al., 2006; Patel et al., 2006; Sutcliffe et al., 2006).
3. Counting (Bayyari & Tudoreanu, 2006; Elmqvist et al., 2006; Qi et al., 2006; Vinayagamoorthy et al., 2006; Wang et al., 2006).

Among other tasks completed less frequently by the participants: estimation, prediction, search and observation.

6.12 Results and Presentation

Another important part of conducting usability studies is the processing and evaluation of the collected data. The processing of the results can be very complex and time consuming since most of the time a lot of data is collected. Therefore it is recommended to employ statistical tools. The most frequently used are: mean, median, frequency distribution, Bonferroni, standard deviation, t-test, and ANOVA (Analysis of Variance).
For the graphical display of the gained data, frequency distributions (in form of histograms) are very popular (83% of the cases in our investigation). Their main purpose is to display error rates and time. As inferential statistics the analysis of variance (ANOVA) is used the most to detect the statistical significance of test results. The ANOVA is a common method of separating the effects of multiple investigation factors (independent variables) on evaluation measures (dependent variables). The ANOVA examines which factors have a significant influence on a dependent variable by comparing the variance within a factor to the variance between factors, (Wanger et al. 1992).
A one-way ANOVA is to be used to estimate the effect of one factor (independent variable) on one of the evaluation measure. A two-way ANOVA is to be used to estimate the effect of two factors (independent variables) on one evaluation measures. According to the literature it is hard to analyze more than two factors using an ANOVA.
In case the effect of a factor is to be estimated on more than one evaluation measure, a multivariate ANOVA (MANOVA) should be applied. A MANOVA is an extension of the ANOVA that reports for multiple dependent variables.
The results of ANOVA's should be displayed in tables, while bar graphs are most used to display descriptive statistics.

7. Case Study

This section briefly presents a case study where the proposed guideline has been applied to. Although it is not imperative to conduct a case study to asses our guideline, we believe this could illustrate its usefulness and application possibilities, and it would facilitate implementation and understanding of the underlying concept.
We have run several case studies over the last few years in order to assess and improve the proposed guideline. The presented case study is related to an innovative use of VR technologies in Telerobotics. The study focuses on a mobile robot teleguide application, (Livatino et al. 2007), which included:

- Qualitative evaluation on user preferences for different VR technologies (Desktop, CAVE, Panorama).
- Quantitative evaluation to support the comparative study, to analyze the advantage of using stereoscopic over monoscopic viewing, and to examine the influence of the user's cognitive profile on his/her performance.

The case study took place at the facilities of Aalborg University in Aalborg (VR Media Lab) and Copenhagen (Medialogy Lab).
The proposed study aimed at improving and extending previous evaluations conducted with different VR facilities, (Livatino & Privitera, 2006). We ran the tests on the three VR facilities represented in figure 1. They are: 3D desktop in mono and stereo; large panoramic wall in mono and stereo; a 6-sided CAVE in stereo only.

7.1 Problem Statement

The problem statement of our case study (part of the test plan, see sub-section 6.9) had three hypotheses. They are:

- Users performing tasks employing stereo visualization perform better than users performing the same tasks employing mono visualization.
- The same task is not performed with the same efficiency and accuracy on different VR facilities.
- The level of visual attention influences the performance of participants in Teleoperation tasks.

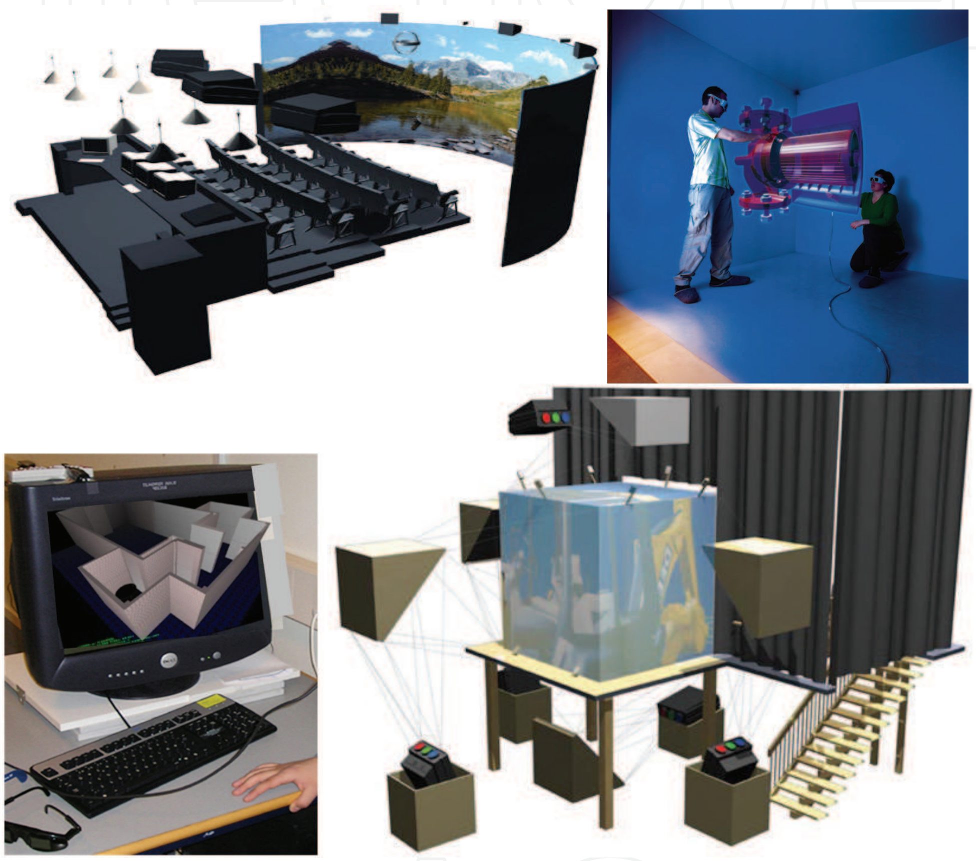


Figure 1. The VR devices tested in the case study are located at the VR Media Lab and Medialogy Lab at the Aalborg University in Aalborg and Copenhagen, Denmark. They are: a 160˚panoramic wall with 8x3 m. screen (top-left), a 3D desktop equipped with shutter glasses (bottom-left), and a 2.5 m^3 6-sided CAVE (top and bottom right). The right figures show the CAVE structure (bottom) and a representative view from inside the CAVE (top)

7.2 Participants

Ten users took part in the evaluation study. The case study required around 2 hours per participant to be completed. All participants had basic to medium experience with VR devices. Experience in playing computer games was also taken into account as experience in VR. Figure 2 shows a user during the evaluation.

7.3 Evaluation Method

As different types of VR devices were compared against each other, the case study was designed as summative evaluation. Because of the limited number of participants and the difficulty of finding equally skilled participants, a within subjects design was preferred over a between subjects design. Therefore each participant fulfilled the same amount of tasks on all available VR devices.

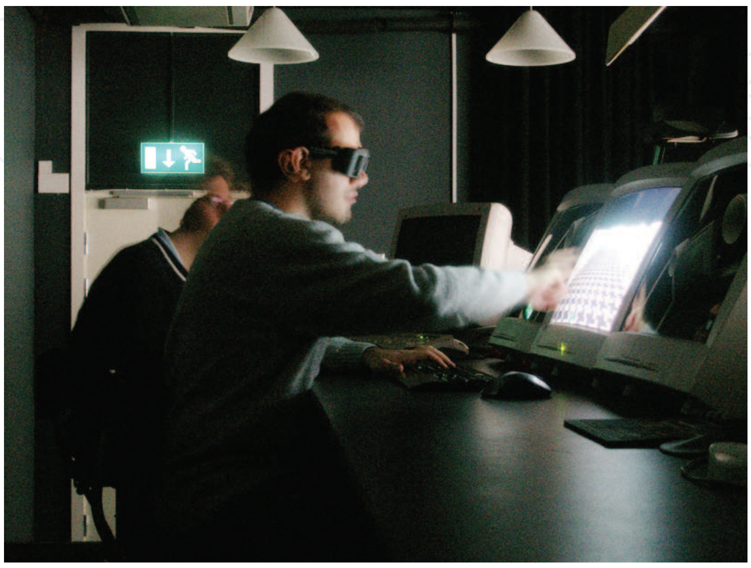


Figure 2. A user during the evaluation on the 3D desktop

7.4 Procedure

The test procedure is part of the test design, (see sub-section 6.9). In our particular case it started with an introduction, then a visual attention test was performed to classify the participants' level of selective visual attention. The users were then asked to teleguide a robot within an interactive test, during which quantitative data were recorded. The last step comprehended the completion of pre-designed questionnaires to acquire qualitative data referring to the users' experience with different VR technologies.

We decided to turn special attention on the counterbalancing of the tasks as well as the sequence during the entire user study to avoid fatigue and learning effects. This required the participants to perform the tests according to a precise schedule.

Practice sessions were administrated before testing. A debriefing phase ended the test session.

7.5 Forms and Questionnaires

As suggested in the guideline, we used a consent form, an information sheet, and different questionnaires for background information and the relation between the different VR devices employed in the test.

During the pilot study the participants were asked to fill in four different questionnaires, one after each task and one at the end of the user study (see Figure 3). The questionnaires contained questions related to users' background, their experience and gaming abilities (e.g. hours per week), specific questions on five proposed judgment categories (adequacy the application, realism, immersion, 3D impression and viewing comfort), and users' overall impression after the user study.

7.6 Evaluation Measures

The following evaluation measures (part of the test plan, see sub-section 6.9), were collected and calculated for the quantitative and qualitative evaluation.

For the quantitative evaluation:

- The numbers of collisions during single driving tasks, (Collision Avoidance test).
- Time to complete each driving task, (Time Completion test).
- Errors made while estimating the relative distance (Access Width Estimation test).
- Number and percentage of tasks completed correctly (Collision Avoidance and Access Width Estimation tests).

For the qualitative evaluation:

- Adequacy of the task to the application, (Adequacy to application).
- The realism of the visual feedback, (Realism).
- Sense of presence, (Immersion).
- Depth impression, (3D impression).
- The user's viewing confort, (Viewing comfort).

The numbers received from the "device questionnaires" were combined with the background and post-test questionnaire.

During the evaluation of the data, the questions were grouped into five categories corresponding to the five qualitative judgement categories, in order to be able to compare the results in each area. The 7 scale semantic differentials were used for the answer of the questionnaires.

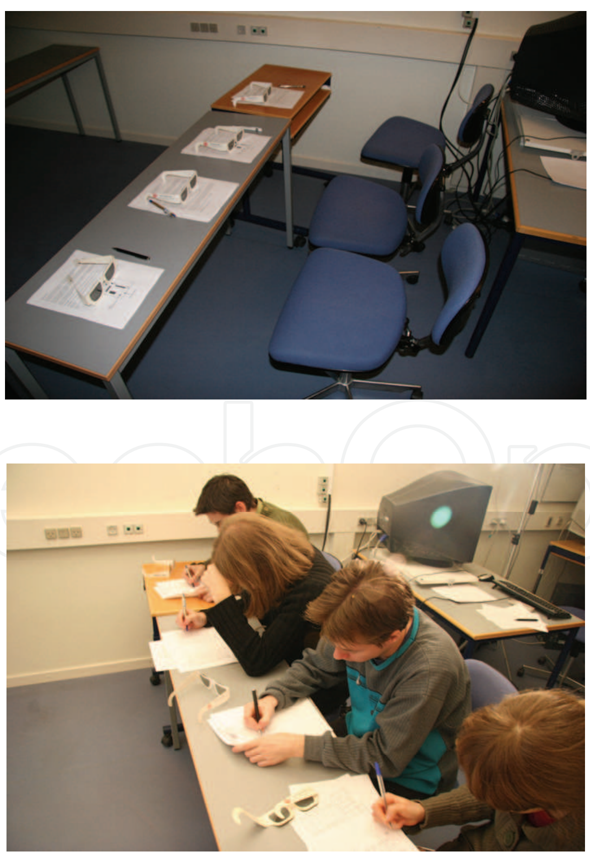


Figure 3. The prepared testing setup including the forms handed to the participants (top) and participants filling in questionnaires (bottom)

7.7 Result Analysis and Discussion

The collected evaluation measures were analyzed through inferential and descriptive statistics and the results were graphically represented by diagrams. In the following some comments on the obtained results are reported while all gathered data can be found in (Koeffel, 2007).

The results on the Panorama and the 3D desktop showed an increased number of collisions with mono visualization compared to stereo visualization.

As for the average completion time needed for the driving task, the participants performed best on the 3D desktop (using either mono or stereo visualization), then on the panorama and CAVE. The participants performed worst on the CAVE. These resulte are shown in figure 4.

Most of the errors in estimating the distance were made using the CAVE. Nevertheless, the CAVE was the facility that users declared to prefer over the others.

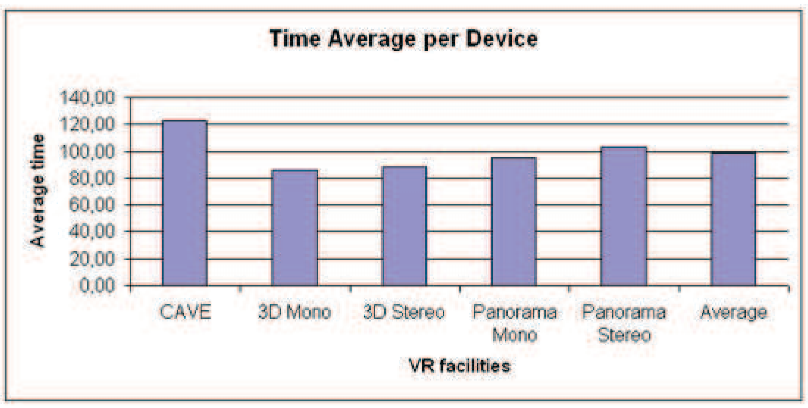


Figure 4. The average completion time per device. As can be seen the users performed slightly faster on the desktop system (mono and stereo), denoted as 3D Mono and 3D Stereo

For what concerns the qualitative evaluation, the CAVE was voted to be most immersive (as expected) while the 3D desktop using mono visualization seemed to be less immersive. The 3D impression and Realism categories emphasized the CAVE as the most appreciated device, while the Viewing comfort was judged without significant differences among all facilities. Figure 5 shows bar diagrams for each judgment category.

When reviewing the overall ranking provided by the post-test and background questionnaire, the CAVE appears again to be the best and most liked device (Figure 6). This goes along with the considerations of Demiralp et al. (Demiralp et al., 2006), telling that "looking-out" tasks (i.e. where the user views the world from inside-out as in our case), require users to use their peripheral vision more than "looking-in" tasks (e.g. small object manipulation). Large and fully surrounding displays also present environment characteristics closer to their real dimension, and in addition the possibility for body interaction allows for a more natural behaviour.

Most of the results of the inferential statistics (ANOVA) did not show significant differences. Though, some clear tendency could be noted. Few significant improvements were found, e.g. the correlation of the number of collisions when using the panorama in stereo instead of mono.

Concerning the analysis related to the influence of the users' selective visual attention, the separation between the single user-classes was based on the median time employed to perform the TAP test. The results showed a trend: users with high-level visual attention performed significantly better in the Collision Avoidance and the Access Width estimation test in the panorama using stereo viewing than with mono visualization (see Figure 7).

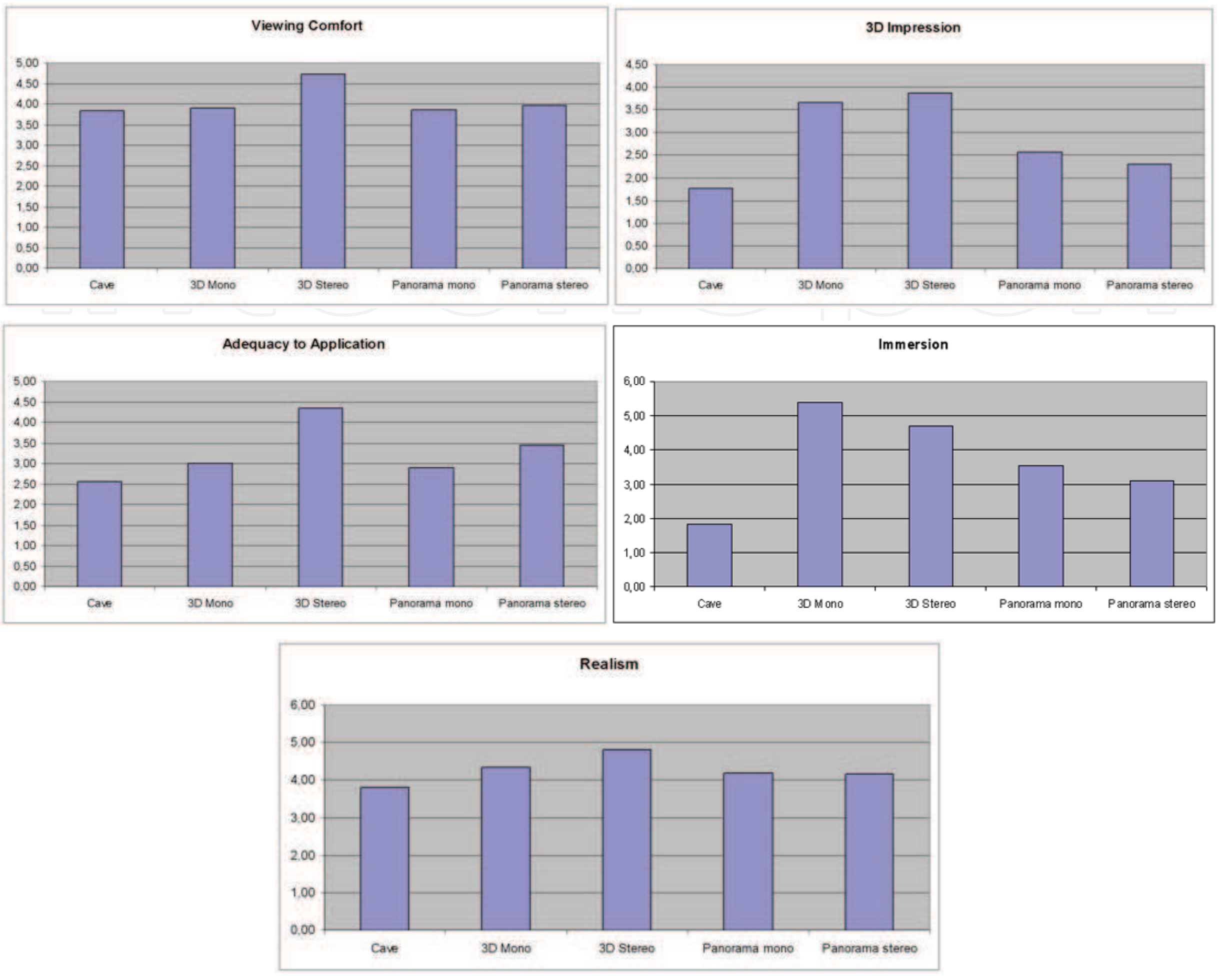


Figure 5. The qualitative results for the categories: Viewing comfort, 3D impression, Adequacy to the application, Immersion and Realism. The 7 scale semantic differentials, ranging from -3 to 3, were converted to a 1 to 7 range for computational purposes

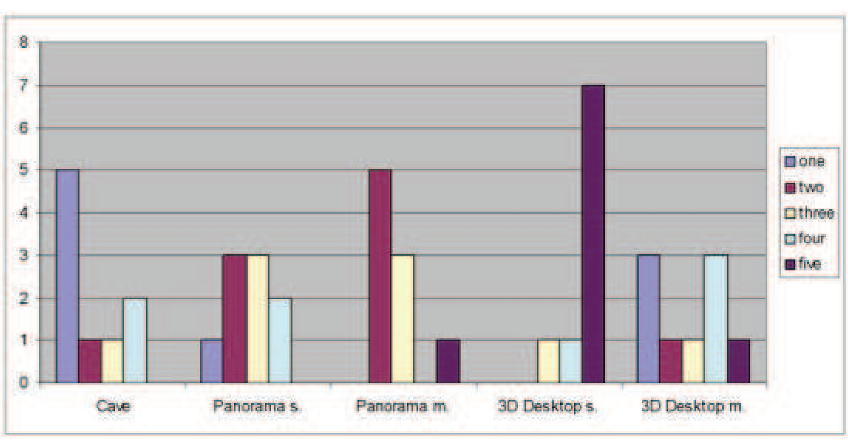


Figure 6. The overall results of the qualitative evaluation indicating that the CAVE was the most appreciated device

Further data processing led to relevant findings. For example, there is a statistically significant difference in the number of collisions between participants that are frequently playing computer games and participants that do not play computer games at all, Figure 8 shows this result.

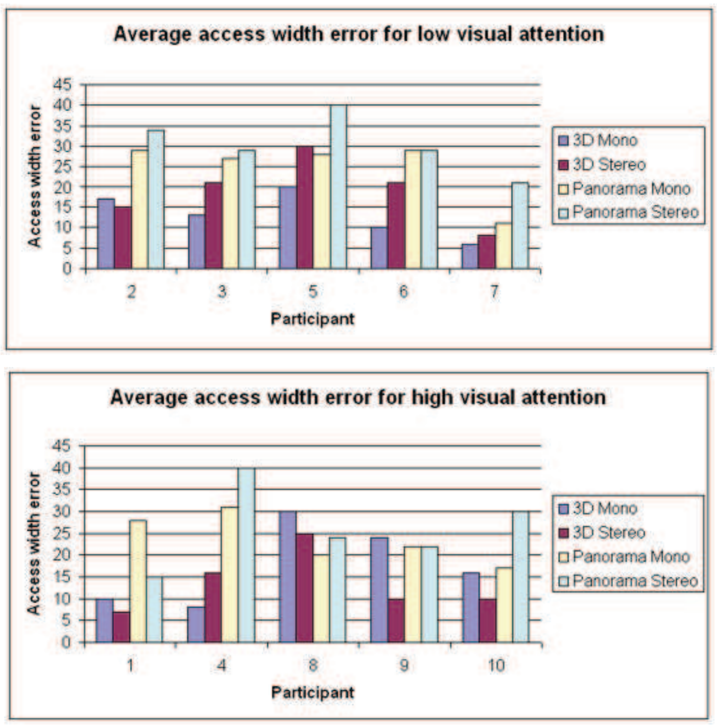


Figure 7. The users' performance of the Access Width estimation on different devices. The top diagram indicates the results for users with a low level of visual attention; the bottom one depicts the results of users with a high level of visual attention

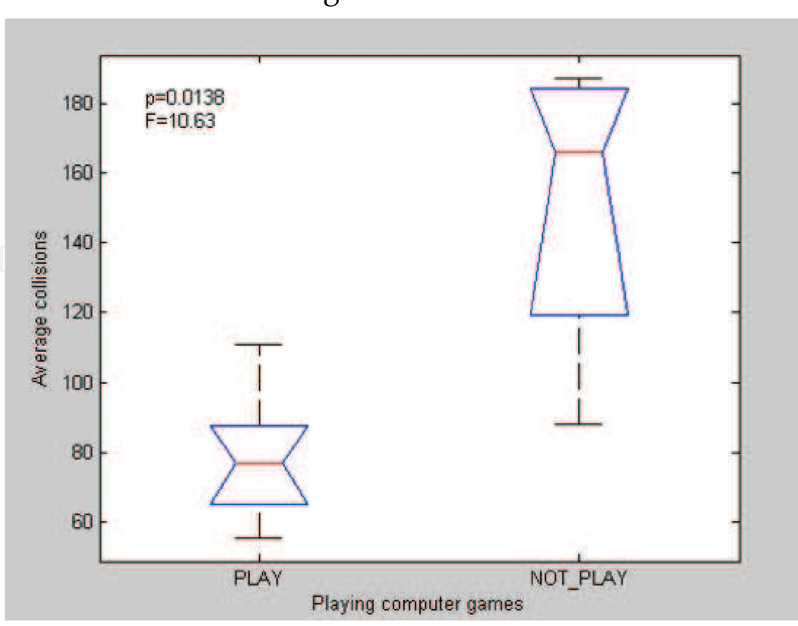


Figure 8. The difference in the users'performance (Collision Avoidance) when playing computer games frequently and when not playing computer games at all, (p=0.0138, F=10.63)

8. Conclusion

The present chapter introduced a guideline for usability evaluation of VR applications. The need for an effort in this direction was underlined in the related literature works. The proposed work targets researchers and students who are not experts in the field of evaluation and usability in general. The guideline is therefore designed to represent a simple set of directives (a handbook) which would assist users drawing up plans and conducting pilot and formal studies.

The chapter briefly introduces multidisciplinary issues related to the evaluation of VR applications and it claims how such guideline may represent a facilitator for multidisciplinary collaborations.

The introduction to the guideline is accompanied by a case study to provide the reader with a practical example of its applicability and to ease its comprehension.

The guideline was judged by users as reasonable and appropriate to the problem. We believe that students and researchers with limited expertise in usability evaluations, as well as those constrained in time, might profit from its content.

Human-Computer Interaction is a subject area in great expansion. There will be therefore an increasing need for user studies and usability evaluations. We believe that a guideline concept as the one proposed will certainly become popular.

9. References

Bach, C. & Scapin, D. L. (2004). Obstacles and perspectives for evaluating mixed reality systems usability.

Bayyari, A. & Tudoreanu, M.E. (2006). The impact of immersive virtual reality displays on the understanding of data visualization. In *VRST '06: Proceedings of the ACM symposium on Virtual reality software and technology*, Limassol, Cyprus, Nov. 2006, pp 368–371.

Bowman, D.A., Gabbard, J.L. & Hix, D. (2002). A survey of usability evaluation in virtual environments: classification and comparison of methods. In *Presence: Teleoperation in . Virtual Environments*, 11(4):404-424

Burdea, G.C., & Coiffet, P. (2003). *Virtual Reality Technology*, John Wiley & Sons, Inc., second edition, ISBN 978-0471360896

Christou, C., Angus, C., Loscos, C., Dettori, A. & Roussou, M. (2006), A versatile large-scale multimodal vr system for cultural heritage visualization. *In Proc. of VRST'06: ACM symposium on Virtual Reality Software and Technology*, Limassol, Cyprus, Nov. 2006, pp.133-140.

Cliburn, D. & Krantz, J. (2008). Towards an effective low-cost virtual reality display system for education. *Journal of Computing. Small Coll.,* 23(3):147–153

Demiralp, C., Jackson, C.D., Karelitz, D.B., Zhang, S. & Laidlaw, D.H. (2006). CAVE and fishtank virtual-reality displays: A qualitative and quantitative comparison. *In proc. of IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 3, (May/June, 2006). pp. 323-330,

Elmqvist, N. & Tudoreanu, M. E. (2006). Evaluating the effectiveness of occlusion reduction techniques for 3D virtual environments. In *proc. of VRST'06: ACM Symposium on Virtual Reality Software and Technology*. Limassol, Cyprus, Nov. 2006.

Faulkner, X. (2000). Usability engineering. Palgrave Macmillan, ISBN 978-0333773215

Fink, P.W., Foo, P.S. & Warren W.H.(2007). Obstacle avoidance during walking in real and virtual environments. *ACM Transaction of Applied Perception.*, 4(1):2

Gabbard, J.L. (1997). A taxonomy of usability characteristics in virtual environments, *Master's Thesis*, Virginia Polytechnic Institute and State University, Blacksburg, Virginia

Gabbard, J. L., Hix, D., & Swan, J. E. (1999). User-centered design and evaluation of virtual environments. *In proc. of IEEE Computer. Graphics. Applications.* 19, 6 (Nov. 1999), 51-59.

Karaseitanidis, I., Amditis, A., Patel, H., Sharples, S., Bekiaris, E., Bullinger, A., & Tromp, J. (2006). Evaluation of virtual reality products and applications from individual, organizational and societal perspectives- the "VIEW" case study. *Int. Journal of Human Perception in Computer Studies* 64, 3 (Mar. 2006), 251-266.

Kasik, D.J., Troy, J.J., Amorosi, S.R., Murray, M.O. & Swamy, S.N. (2002). Evaluating graphics displays for complex 3D models. *IEEE Computer Graphics and Applications*,vol.22, no.3, (May/Jun 2002) pp.56-64

Koeffel, C. (2007). Evaluation methods for virtual reality applications, *Tech.Rep*. Medialogy, Aalborg University, Denmark, 2007.

Koeffel, C. (2008). Handbook for evaluation studies in vr for non-experts, *Tech.Rep*. Medialogy, Aalborg University, Denmark, 2008.

Livatino, S. & Koeffel, C. (2007), Handbook for evaluation studies in virtual reality. *In proc. of VECIMS '07: IEEE Int. Conference in Virtual Environments, Human-Computer Interface and Measurement Systems,*, Ostuni, Italy, 2007

Livatino, S., Gambin, T., Mosiej L, & Koeffel, C. (2007). Exploring critial aspects in vr-based mobile robot teleguide. *In proc. of RA'2007: Robotics and Applications conference*, Wurzburg, Germany, 2008

Livatino, S. & Privitera, F. (2006). Stereo visualization and virtual reality displays. *In proc. of Vis2006: IEEE Visualization 2006 conference*, Baltimore, USA, November, 2006.

Malkondu, E. (2007). The study of the camera parameters in remote robot teledriving. *Master's thesis*, Aalborg University, Denmark, 2007.

Marsh, T. 1999. Evaluation of virtual reality systems for usability. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, May 15 - 20, 1999). CHI '99. ACM, New York, NY, 61-62.

McMahan, R.P., Gorton, D., Gresock, J., McConnell, W. & Bowman, D.A. (2006). Separating the effects of level of immersion and 3d interaction techniques. *In proc. of VRST '06: ACM symposium on Virtual Reality Software and Technology*, Limassol, Cyprus, Nov. 2006 pp. 108–111

Ni, T., Bowman, D. A., & Chen, J. (2006). Increased display size and resolution improve task performance in information-rich virtual environments. *In Proc. of Graphics Interface 2006*, Quebec, Canada, June 07 - 09, 2006. (ACM Int. Conference Proceeding Series, vol. 137. Canadian Information Processing Society, Toronto, Ont., Canada, 139-146).

Nielsen, J. (1993). *Usability engineering*, Morgan Kaufmann, ISBN 978-0125184069

Nielsen, J., & Mack R.L. (1994). *Usability Inspection Methods*, John Wiley & Sons, New York, USA, May 1994, ISBN 978-0471018773

Patel, H., Stefani, O., Sharples, S., Hoffmann, H., Karaseitanidis, I. & Amditis, A.(2006). Human centred design of 3-d interaction devices to control virtual environments. *Int. Journal of. Human Perception in Computer Studis.*, 64(3):207–220

Qi, W., Russell, I., Taylor, M., Healey, C.G. & Martens, J.B. (2006). A comparison of immersive hmd, fish tank vr and fish tank with haptics displays for volume visualization. *In proc. APGV '06: 3rd Symposium on Applied Perception in Graphics and Visualization*, New York, NY, USA, pages 51–58

Rubin, J. (1994). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons, ISBN 978-0471594031

Sarodnick, F., & Brau, H. (2006). *Methoden der Usability Evaluation, Wissenschaftliche Grundlagen und praktische Anwendung*, Hans Huber Verlag, ISBN 978-3456842004

Sutcliffe, A. & Gault, B. (2004). Heuristic evaluation of virtual reality applications, *Interacting with Computers*, Volume 16, Issue 4, (August 2004) Pages 831-849

Sutcliffe, A., Gault, B., Fernando, T. & Tan, K. (2006). Investigating interaction in cave virtual environments. *ACM Transaction Compute.-HumanInteraction*, 13(2):235–267

Takatalo, J., Nyman, G. & Laaksonen, L.(2008). Components of human experience in virtual environments. *Computer Human Behaviour*., 24(1):1–15

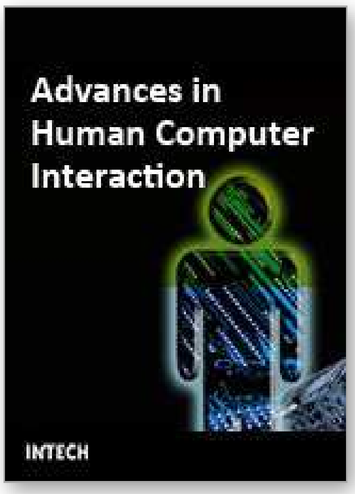
Tromp, J.G. & Nichols, S.C. (2003). VIEW-IT: a vr/cad inspection tool for use in industry. *In proc. of the HCI International 2003 Conference*, Crete, 22-27 June.

Vinayagamoorthy, V., Brogni, A., Steed, A. & Slater, M.(2006). The role of posture in the communication of affect in an immersive virtual environment. *In proc. VRCIA '06:ACM international conference on Virtual reality continuum and its applications*, New York, NY, USA, 2006, pp 229–236

Wanger, L.R., Ferweda J.A., Greenberg, D.P. (1992). Perceiving spatial reletionships in computer generated images. *In Proc. of IEEE Computer Graphics and Animation.*

Wang, Y., Otitoju, K., Liu, T., Kim, S., & Bowman, D. A. (2006). Evaluating the effects of real world distraction on user performance in virtual environments. *In Proc. of VRST'06: ACM Symposium on Virtual Reality Software and Technology.* Limassol, Cyprus, Nov. 2006.

Wilcox, L.M., Allison, R.S., Elfassy, S. & Grelik, C. (2006). Personal space in virtual reality. *ACM Transaction of. Applied Perception*, 3(4):412-428.



Advances in Human Computer Interaction
Edited by Shane Pinder

ISBN 978-953-7619-15-2
Hard cover, 600 pages
Publisher InTech
Published online 01, October, 2008
Published in print edition October, 2008