TOWARDS AN EFFICIENT METHODOLOGY FOR EVALUATION OF QUALITY OF EXPERIENCE IN AUGMENTED REALITY

Jordi Puig[1], Andrew Perkis[1], Frank Lindseth[1,2], Touradj Ebrahimi[3]

Centre for Quantifiable Quality of Service in Communication Systems (Q2S)[1]
Norwegian University of Science and Technology (NTNU), Trondheim, Norway, SINTEF[2], EPFL[3]

ABSTRACT

The goal of this paper is to survey existing quality assessment methodologies for Augmented Reality (AR) visualization and to introduce a methodology for subjective quality assessment. Methodologies to assess the quality of AR systems have existed since these technologies appeared. The existing methodologies typically take an approach from the fields they are used in, such as ergonomics, usability, psychophysics or ethnography. Each field utilizes different methods, looking at different aspects of AR quality such as physical limitations, tracking loss or jitter, perceptual issues or feedback issues, just to name a few. AR systems are complex experiences, involving a mix of user interaction, visual perception, audio, haptic or other types of multimodal interactions as well. This paper focuses on the quality assessment of AR visualization, with a special interest on applications for neuronavigation.

Index Terms— Quality of Experience, Augmented Reality, Subjective Quality Assessment, Objective Measurement.

1. INTRODUCTION

Joint efforts have been made to standardize quality measurement for audiovisual communications. Quality of experience (QoE) is not well defined although efforts are underway to better understand its meaning and mechanisms. The prevailing definition is often referred to that by the International Telecommunication Union (ITU) as: "The overall acceptability of an application or service, as perceived subjectively by the end-user". This is well known and much used in the audio-visual communications field from an engineering perspective and is usually applied to assessing audio or video perception. Their methods comprehend objective quality metrics and subjective quality assessment tests. The ITU has described non-interactive subjective assessment methods for evaluating the one-way overall video quality for multimedia applications in ITU P.910. This work has allowed important advances for the media industry.

However, most of the work regarding the assessment of perception of multimedia systems has focused on individual modalities, i.e., audio and video separately. Although important work has recently been performed on perceptual-based audio–visual quality metrics [1], and a taxonomy of QoE for the assessment of multimodal human-machine interaction has been defined in [2], it seems that this evaluation methods are still very far to fulfill the current needs for AR assessment. Current assessment methods seem to be not applicable to AR systems since they usually assume the end user as a passive entity. AR systems are based on interaction and more importantly in an active perception and experience of the content.

Augmented Reality (AR) refers to the addition of a computer-assisted contextual layer of information over the real world, creating a reality that is enhanced or augmented. Azuma [3] defines Augmented Reality (AR) as systems that have the following three characteristics:

1) Combine real and virtual
2) Interactive in real time
3) Registered in 3D

It is an obvious fact that AR has been traditionally related to visual feedback. Milgram describes in 1994 "A taxonomy of mixed reality visual displays" [4] giving the first insights to differentiate the relations between the real and the virtual in the well known "virtuality continuum". Furthermore he sets several dimensions to differentiate display types i.e. "Extent of World knowledge", "Reproduction Fidelity" and "Extent of Presence Metaphor". Therefore every display type has different implications and consequences on the user experience. At this point "see-through" displays like head-mounted displays (HMD's) and "monitor based" displays are discussed and categorized depending on their use.

A decade later Bimber [5] presents an extensive work on displays for "Spatial augmented reality" where many other types of displays are discussed, e.g. "projector based",

[1] "Centre for Quantifiable Quality in Communication Systems, Centre of Excellence" appointed by the Research Council of Norway, funded by the Research Council, NTNU and UNINETT. http://www.q2s.ntnu.no/

"optical overlays", "retinal displays", "auto-stereoscopic displays", etc. Thus there are many different ways of presenting virtual information, and some might constitute a better solution depending on their actual use. We will understand a display as something very heterogeneous, frameless, very far from the dogmatic conception of a limited square enclosing content.

There are two major issues related to visual quality in AR. At first, perceptual issues can be addressed depending on the characteristics of the display devices. Secondly, closely related but clearly of a different kind, are the perceptual issues derived from semiotics and visual design issues of the rendered application.

Issues of the first group were early addressed by Drascic & Milgram in [6] where problems like depth cues e.g. pictorial depth, kinetic depth, physiological depth or binocular disparity are discussed. Implementation errors e.g. calibration errors, calibration mismatches and inter-pupillary distance mismatches and technological limitations e.g. dynamic registration mismatches, restricted field of view, resolution, luminance and contrast mismatches amongst others. The list on perceptual issues on device engineering is long and has been recently revised and extended in [7]. All those problems do not take into account the aspects related to visual design of the graphical user interfaces.

There are visualization aspects in AR being approached from a designer's perspective. Leaving aside the concerns on the display technologies used, one can solve the need for a visual feedback using a number of different metaphors. Strategies like masking, zooming, highlighting, or offering different levels of visual information load can be highly determining on the final quality of an AR system. Examples of these solutions have been shown in [8] and [9]. To summarize this point, there are different levels of quality for an AR system, from the more tangible aspects of device physical properties to the visual aspects of the virtual information displayed. All of them approached and evaluated from distinct perspectives depending on their focus.

AR technologies have been introduced in the medical field over the last decade [10], and continue to advance the field [11]. The discussion on which AR systems have a better performance in the laparoscopic field has been discussed in [12] with the conclusion of a lack of a consistent assessment protocols for such technologies. Another major work is focusing on the current AR visualization technologies [13]. However, there is a lack of focus on the assessment methods for their evaluation.

The set of computer-assisted technologies used for the treatment of neuronal injuries are denoted as neuronaviation systems. Nowadays surgeons can use AR neuronavigation during pre-operative planning to assess the properties of a lesion. Such systems demand a high visual and interactive quality to ensure successful operations.

Still today there is no consensus on how to assess the quality across different visualization aspects of AR. In section 2 we discuss how scientific disciplines have tackled distinct assessment methodologies with different consequences. Section 3 discusses the differences between the fields. Section 4 gives an introduction on the need for assessment methods in neuronavigation. Section 5 focuses on the special requirements of assessment methods in AR visualization. Section 6 proposes a methodology derived from the existing methods from the fields of ergonomics, usability and Quality of Experience (QoE) to approach the question of quality assessment in AR visualizations. And section 7 is a proof of concept on the method. Finally, section 8 summarizes the conclusions of this research.

2. ASPECTS OF QOE IN AR

Human Computer Interaction (HCI) is traditionally assessed according to usability and ergonomics, which are characterized as human factors. The other disciplines working on this context are closely related, but they refer to different aspects when describing the quality of their system.

2.1. Usability

Although usability is an interdisciplinary concept, it is understood as the field concerned with the usage of a system. Usefulness is a concept typically composed by factors such as learnability, efficiency, memorability, satisfaction and errors [14]. A usability analyst assesses the mentioned aspects separately and the outcome is applied to a product or application.

2.2. Ergonomics

Even though ergonomics appears to be similar to usability, it has had a very different history and development. The word was born in the Ancient Greece and could be translated to "natural laws in work" which points to the need of finding best practices at work. Today it is most commonly used in industrial design and basically considers the relation of the objects to the human body. Ergonomic studies aim at reducing body strain injuries and to optimize the forms of the objects to economize body movements amongst others. A major work relating ergonomic quality of interactive systems has been described in [15].

2.3. Human Factors

Human factors could be understood as the interdisciplinary field that comprehends all aspects for the study of HCI

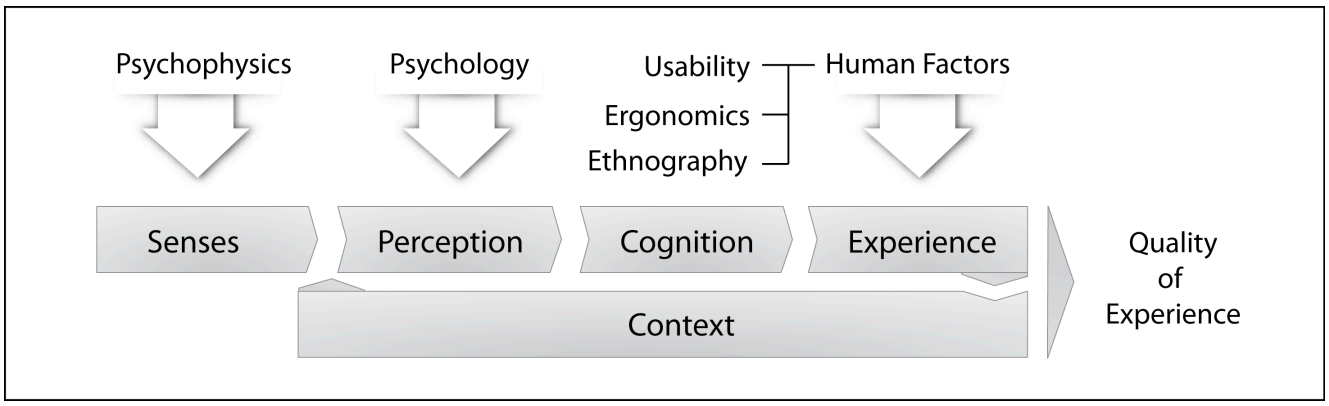

Figure 1. Quality of a stimuli assessed at a different stages and the target fields of study.

technologies. In HCI, heuristic evaluation is a usability-testing technique carried out by expert usability consultants by using guidelines, checklists, standards, etc. The evaluations are carried out in real environments, saving time resources. This is known as soft criteria, because it relies on the experience of the evaluator.

2.4. Ethnography

The case of ethnographical assessment is slightly different. The main focus is set on the cultural factors influencing HCI. The basis is that humans have strong cultural interactions, hence the main goal is to evaluate how well a system supports the knowledge and activities of a group of people. These assessments are especially relevant when the evaluated system is used by groups of people.

2.5. Subjective Quality Assessment

Subjective assessments can be used to evaluate a given quality, when putting the user as the focal point. It does not rely on the eyes of an experienced evaluator. Instead it uses large amounts of user data and statistical processing. When "statistical significances" emerge from the data, scientific conclusions are drawn. These methods are very reliable when applied to perceptual measurements where cognitive interactions are minimized. When subjective assessments are carried out, the subject gives answers to questions about the quality of the stimulus. This fact assumes that the user can consciously give a correct and accurate answer to what is being asked. Another particularity is that the tests are carried out in laboratories, which means that external variables are intentionally minimized. These methodologies do not rely on the experience of an evaluator but rather rely on the self-assessment of each individual subject, which is processed by statistical means to extract a reliable conclusion. Quality assessment in the fields of signal processing, food science and acoustics are typically using such methods.

2.6. Psychophysics

The methods in psychophysics are of harder criteria. Usually there is no self-assessment; the cognition of the subject is not represented in the measurements. The aim is to quantitatively assess the relationship between physical stimuli, which can be an image or a sound, and the perceptions they create. In this case the tests are also carried out in a laboratory isolating the subject from many external parameters present in the real world.

3. THE DIFFERENCES BETWEEN FIELDS

We could intuitively assume that the so called "inspection methods" or "qualitative methods" performed by ethnographers, and Human Factors scientists are wider. They can target multiple issues at the same time but they have softer empirical grounds of truth. Furthermore, QoE is narrowing the chances for wrong conclusions by narrowing also the assessed variables and the undesired noise from reality. But the main difference between both methods is that QoE has rarely been used in interactive systems while in human factors the assessments always focusing on human-computer interaction.

The relation between the quality of a stimuli assessed at different stages and the target fields of study is depicted in Fig. 1. While Pereira [15] proposes a triple sensation-perception-emotion user model for approaching multimedia experience, we envision a direction with five stages starting from the senses of the subjects when receiving a stimulus. At this stage quantitative psychophysical methods may be applied to assess them. From the senses more complex perceptions may develop, leading to the second stage. Subjective quality evaluation is the assessment method to evaluate the perceived quality of the stimuli. While the understood stimuli interact with the attention processes and memory we enter in the cognitive domain, also being under the field of QoE. Closely related and in a higher level the cultural effects interact with the cognitive processes. At this point ethnographical methods are suitable to assess such

cultural influences. At the end of the quality process the final experience is determined. At this point heuristic evaluation is used to study the overall quality of a multimedia experience. Finally, we understand the context as the phenomena existing from perception to experience.

4. THE CASE OF NEURONAVIGAITON

The need for evaluation methods is tangible since several technologies are already being commonly used. The medical field has adopted AR technologies since their appearance and nowadays they are already utilized on an everyday basis.

In surgical interventions surgeons rely on imaging technologies during the entire process, and AR systems are increasingly being used. This is especially true in the planning phase right before the procedure starts. A major challenge during the intervention is the fact that AR systems are not well adapted to the surgical workflow. In addition few studies have been conducted to assess the benefits of AR systems compared to more traditional visualization techniques. The result is that only enthusiasts use intra-operative AR systems and most surgeons resolve slicing through the image volumes. AR would probably be more useful bridging the gap between the information from modalities like MR and CT in one side and real-time 2D data from endoscopes, microscopes and ultrasound on the other side. Today this information is often found on separate displays and it will be crucial to merge all the information in the future. In order to achieve this it's important that AR equipment is adapted to the surgical field and to have methods to assess their quality.

Within the national center for ultrasound and image guided therapy in Trondheim, Norway (a collaboration between SINTEF, NTNU and St Olavs University hospital), surgeons and engineers have been working closely together for over 15 years in order to advance the neuronavigation technology. This has been done by letting engineers participate in the operating room and by conducting qualitative analysis and informal interviews. Current research has disclosed an increasing need for quantitative assessment methods to assure quality for the systems currently under development.

5. QUALITY ASSESSMENT FOR AR VISUALIZATIONS

As previously stated, AR systems are complex. Interactivity, haptics, virtual visual and auditory perception merged with an enacted perception [16] of reality. Quality assessment is still required and nowadays becoming crucial. AR assessments cannot escape from the fact that evaluations must include interaction. Therefore quantitative methods must adapt to this context. Ergonomic scientists have usually assessed AR systems by offering users to perform a task to reach a goal [17]. However, most of these studies have been focusing in the overall performance of the subject. We want to focus on quality metrics for the visual perception on AR. The fact that AR systems clearly make use of action during the perceptual process is what pushes us to present a task dependent methodology to quantitatively evaluate AR visualizations.

6. THE PROPOSED METHOD

Our proposed methodology uses subjective assessment and objective measurements. The subjective measurements can be in the form of questionnaires, subjective user ratings or judgments. Objective measurements are processed through a task to be accomplished by the subject. The task is carried out using a tracked device, which allows the recording of the user interaction in an accurate manner. The performance of the user is analyzed by processing time to completion (TTC), accuracy or error rates by statistical means. By running a sufficient amount of subjects through two experimental conditions, i.e. visualization A versus visualization B, it will be possible to conclude whether there is a statistical difference between the performance of users. These conclusions will be combined with the outcome of the analyzed scores of the user ratings. This method can be carried out in a laboratory with non-expert users.
In case the results of the experiment are conclusive further evaluations can be conducted in the real environment to achieve deeper understanding of the future needs for the system. At this stage a qualitative assessment is recommended to improve any aspects related to the efficiency of a given technology in the real environment.

7. A PROOF OF CONCEPT

We conducted a pilot experiment following the proposed method but concentrating only in objective measurements. Subjective assessments have to be tailored to specific experimental designs; therefore such measurements are not suitable for a pilot experiment. A full implementation of the proposed method is to be detailed in future publications.
The experiment consisted in the comparison of two different presentation types i.e. a computer screen and a projection of a bigger size. The task to accomplish consisted in matching the position and orientation of a virtual object (Fig. 2 in red) by manipulating a tracked marker with an augmented similar object (Fig.2 in white). Although this task does not emulate a realistic neuronavigation system (basically because it does not use the same visualization system), it is similar in terms ergonomics, interaction and human behaviour. The subjects have to coordinate the visual feedback on the presentation with their movements in order to precisely match the position on the real space. The task was repeated sequentially using the same presentation type.

Figure 2. Matching position and orientation task for a pilot experiment.

A total of 10 different positions had to be completed. Once the first presentation type was assessed, the same task was completed in the other presentation type.

Subjects 1 and 3 started with presentation type A and subjects 2 and 4 with type B. For the pilot experiment TTC was measured for each position. The improvement in the learning curve of the subjects was more evident in presentation type B compared to type A (see Fig. 3 and 4). Subjects required some minutes to match the first targets and reduced the times to some seconds at the end of the measurements. The proof of concept is promising and will be used as a guide to refine and optimize our methodology.

Preliminary studies show that this method can offer successful results in objective evaluations by giving evidence of the performance in compared presentation types. On the other hand, subjective ratings and questionnaires can better explain the source of performance rates obtained.

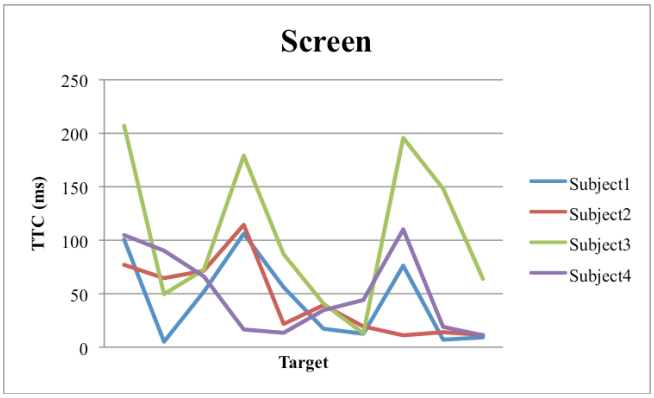

Figure 3 TTC Scores for presentation type A

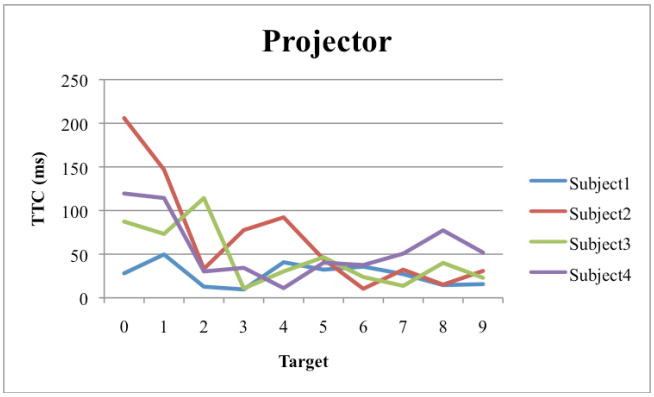

Figure 4 TTC Scores for presentation type B

8. CONCLUSIONS

We exposed the issues and disciplines related to quality assessment methodologies for AR visualizations. We proposed a quality assessment method for AR visualizations. The method is based on quantitative evaluation with a mixed approach using subjective assessment and objective measurements. The intention of this research is to apply the method to assess technologies in the neuronavigation field.

9. REFERENCES

[1] J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio–visual services: A survey," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 482–501, Aug. 2010.

[2] S. Moller, K.-P. Engelbrecht, C. Kuhnel, I. Wechsung, and B. Weiss, "A taxonomy of quality of service and Quality of Experience of multimodal human-machine interaction," *Quality of Multimedia Experience, 2009*. QoMEX 2009. International Workshop on, pp. 7–12, 2009.

[3] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators & Virtual Environments*, vol. 6, pp. 355–385, 1997.

[4] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," IEICE *Transactions on Information and Systems*, vol. 77, no. 12, pp. 1321–1329, 1994.

[5] O. Bimber and R. Raskar, *Spatial Augmented Reality: Merging Real and Virtual Worlds*. A K Peters, Ltd., 2005, p. 392.

[6] D. Drascic and P. Milgram, "Perceptual issues in augmented reality," *PROCEEDINGS-SPIE THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING*, pp. 123–134, 1996.

[7] E. Kruijff, J. Swan, and S. Feiner, "Perceptual issues in augmented reality revisited," *9th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2010, pp. 3–12, 2010.

[8] D. Kalkofen, M. Tatzgern, and D. Schmalstieg, “Explosion Diagrams in Augmented Reality,” in *Virtual Reality Conference*, 2009. VR 2009. IEEE, 2009, pp. 71–78.

[9] E. Mendez and D. Schmalstieg, “Importance masks for revealing occluded objects in augmented reality,” *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, pp. 247–248, 2009.

[10] J. H. Shuhaiber, “Augmented reality in surgery,” *Archives of Surgery*, vol. 139, no. 2, p. 170, 2004.

[11] P. Lamata, W. Ali, A. Cano, J. Cornella, J. Declerck, O. J. Elle, A. Freudenthal, H. Furtado, D. Kalkofen, and E. Naerum, “Augmented Reality for Minimally Invasive Surgery: Overview and Some Recent Advances,” *Augmented Reality*, pp. 978–953, 2010.

[12] S. M. B. I. Botden and J. J. Jakimowicz, “What is going on in augmented reality simulation in laparoscopic surgery?,” *Surg Endosc*, vol. 23, no. 8, pp. 1693–1700, Sep. 2008.

[13] T. Sielhorst, M. Feuerstein, and N. Navab, “Advanced Medical Displays: A Literature Review of Augmented Reality,” *Journal of Display Technology*, vol. 4, no. 4, pp. 451–467.

[14] Nielsen, J. (1994) *Usability Engineering*, Morgan Kaufmann Publishers, ISBN 0-12-518406-9.

[15] D. L. Scapin and J. M. C. Bastien, “Ergonomic criteria for evaluating the ergonomic quality of interactive systems,” *Behaviour & Information Technology*, vol. 16, no. 4, pp. 220–231, Jan. 1997.

[16] F. Pereira, “Sensations, perceptions and emotions: towards quality of experience evaluation for consumer electronics video adaptations,” *Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005.

[17] A. Noë, *Action in perception*. The MIT Press, 2004, p. 277.

[18] A. Tang, C. Owen, F. Biocca, and W. Mou, “Comparative effectiveness of augmented reality in object assembly,” *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 73–80, 2003.