Quality of Experience for a Virtual Reality simulator

Kjell Brunnström[a,b], Mårten Sjöström[b], Muhammad Imran[c,b], Magnus Pettersson[c], Mathias Johanson[d]
[a]RISE Acreo AB, Kista, Sweden
[b]Mid Sweden University, Sundsvall, Sweden
[c]HIAB, Hudiksvall, Sweden
[d]Alkit Communications AB, Mölndal, Sweden

Abstract

In this study, we investigate a VR simulator of a forestry crane used for loading logs onto a truck, mainly looking at Quality of Experience (QoE) aspects that may be relevant for task completion, but also whether there are any discomfort related symptoms experienced during task execution. The QoE test has been designed to capture both the general subjective experience of using the simulator and to study task completion rate. Moreover, a specific focus has been to study the effects of latency on the subjective experience, with regards both to delays in the crane control interface as well as lag in the visual scene rendering in the head mounted display (HMD). Two larger formal subjective studies have been performed: one with the VR-system as it is and one where we have added controlled delay to the display update and to the joystick signals. The baseline study shows that most people are more or less happy with the VR-system and that it does not have strong effects on any symptoms as listed in the Simulator Sickness Questionnaire (SSQ). In the delay study we found significant effects on Comfort Quality and Immersion Quality for higher Display delay (30 ms), but very small impact of joystick delay. Furthermore, the Display delay had strong influence on the symptoms in the SSQ, and causing test subjects to decide not to continue with the complete experiments. We found that this was especially connected to the longer added Display delays (≥ 20 ms).

Introduction

Virtual and augmented reality (VR, AR) are emerging technologies for assisting or solving real world industrial problems. We consider in this case immersive techniques, where the user is visually interacting with the physical environment using Head-Mounted Displays (HMD), also popularly denoted "VR goggles". Potentially this will imply that workers will be using such goggles for extended periods of time; not only the same day, but most likely each working day for a long time. Therefore, the quality related issues are crucial, not only because they are tied to performance and task completion, but also because of the well-being of the worker.

In this study, we investigate a VR simulator of a forestry crane used for loading logs onto a truck, mainly looking at Quality of Experience (QoE) aspects that may be relevant for task completion, but also whether there are any discomfort related symptoms experienced during task execution. The target system is an immersive video based system with the ambition to also become an AR system that lets the crane operator stay in the truck cabin while loading logs onto the truck, aided by a 270° HMD video view generated from four video cameras mounted on the crane (see Figure 1). The benefits of this system are that the crane does not need to be equipped with an operator cabin as well as improved safety and comfort for the operator. Connected to the development of the system, a desktop simulator has also been developed (see Figure 2), which instead of the live video views generates a virtual view using a 3D gaming engine. The VR simulator is used as an educational tool and should simulate as closely as possible the actual crane system. The present QoE study has focused on the VR simulator, with the intention to be a starting point for assessing the subjective experience also of the AR system. Both the AR system and the VR simulator has the same crane control devices (joysticks) as the real truck cabin and an Oculus Rift HMD for the visual information.

The QoE test has been designed to capture both the general subjective experience of using the simulator and to study task completion rate. Moreover, a specific focus has been to study the effects of latency on the subjective experience, with regards both to delays in the crane control interface as well as lag in the visual scene rendering in the HMD. The reason latency is of particular interest is twofold: Firstly, it is a crucial design parameter for the AR system, since the processing of video signals to generate the visual HDM scene is very CPU consuming and the tolerable delay hence serves as a performance requirement for the processing hardware of the system. Secondly, we are interested in exploring the possibility of controlling a crane from a remote location, which requires the video signals as well as the crane control signals to be transmitted over a (typically wireless) network connection, which will introduce delays, and the delay tolerance hence strongly influences the feasibility of such an approach.

Figure 1: Picture of VR-goggle based crane operation from inside the Truck cabin

Background

Augmented Telepresence

To highlight the focus and the direction of our work we are using the term Augmented Telepresence (AT) to denote applications where high-quality video-mediated communication is the enabling technology, but where additional data can be superimposed or merged with the video as in Augmented Reality. It is not yet a commonly used term, but has previously been used by a few authors [1, 2].

AT is similar to augmented reality in that it tries to present additional information on top of the view seen by the user. It differs from augmented reality in that primarily the user is present in a remote location seeing the augmented view, but may also include the case were a two-way audio and/or audio-visual communication is being retained at the same time with the user seeing the augmented view.

Quality of Experience

Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in light of the user's personality and current state, as defined by EU Cost Action 1003 Qualinet [3]. Although this version of the definition of QoE is not yet standardized, it is supported by large number of scientist in the field and, most likely, the standards will be updated to follow this definition. A comprehensive overview of the field can be found in the recent QoE book by Möller and Raake [4].

The above definition of QoE, which is also pointed out by Möller and Raake [4], goes beyond the traditional QoE and Quality of Service (QoS) research and then makes a clear overlap with the User Experience (UX) research tradition. These two fields come from a clearly different research tradition and community i.e. Telecommunication and Human Computer Interaction respectively. The QoE community is still in the process of embracing some of the more user centric and UX like methods.

Traditionally, in the QoE research, the methods to gain insight into the delivered quality of a service and the users' experience of it have been done through controlled laboratory experiments, where the opinions of panels of users have been collected. The results are reported in Mean Opinion Scores (MOS). These methods are very often referred to as subjective quality assessment methods and there are standardized ways of conducting them e.g. for visual quality, ITU-R Rec. BT.500-13[5] or ITU-T Rec. P.910[6]. These methods have been criticized for not providing enough ecological validity [7]. Improvements have been done for example in ITU-T Rec. P.913[8]. The intensive investigations into 3D video quality a few years ago, when the 3D TV hype was the most intense, have now resulted in new Recommendations from the ITU [9-11]. It was discovered that if care was not taken, several user experience issues such as discomfort and visual fatigue may occur. The Recommendations give some guidance on how to minimize these. An attempt to build an experimental framework for QoE of AR was made by Puig et al. [12] who advocate a combination of subjective assessment (e.g. questionnaires, subjective ratings) and objective measurements (e.g. task completion time, error rates). They only presented the results from a pilot study, so it still needs to be experimentally confirmed whether the framework gives scientifically reproducible results and if it can be extended to AT.

Now we are in the early stages of large scale deployment of fully immersive environments e.g. Oculus Rift, PS4 VR, or HTC Vive. Furthermore, the development of 5G will give higher bandwidth, and more importantly, low latency mobile networks. This means that we are now facing low latency immersive environments on a large scale, meaning that it is of utmost importance to understand the user experience issues connected to it. New types of interaction, especially those of a highly immersive nature, will put new demands on the correct way of designing the user environment. Therefore, increased efforts should be allocated to understanding the QoE for not inducing a negative perceived user experience, discomfort or even simulator sickness. Furthermore, low latency can enable services and applications with an intensive interaction component such as gaming or remote control of professional equipment, which will increase the load on the user. Although such research has been ongoing for some time, the technical development and increasing availability of immersive low latency user environments make research more urgent.

Related work

This section presents some related work that deals with measuring quality of experience of VR-simulators in different perspective and involving visual and/or haptic delay.

Debattista et al. [13] presented a subjective evaluation of high fidelity virtual environments for driving simulations. The evaluation was based on 44 participants; providing them access to real world and purposely build representative virtual environment with graphics quality settings of low, medium and high. The study concluded that graphics quality affects the perceived fidelity of visual and overall experience. However, the study was limited to only fixing graphics quality in three fixed states and the author acknowledge the complexity of visual simulator.

Strazdins et al. [14] studied virtual reality in the context of gesture recognition for deck operation training. Since existing simulators used only keyboards and joysticks for getting input, the authors developed a prototype of gesture recognition system and performed study on 15 people. The study concluded that improving video quality affects the user experience positively; better quality improved score from 3.63 to 4.83 on a scale of 5. However, participants' self-assessment score; measuring how well they performed was 3.7 on 5 scale. It is worth mentioning that the study was performed on students with no actual crane operation experience.

Suznjevic et al. [15] compared the QoE of two different VR-goggle technologies i.e. Oculus Rift and HTC Wive in a pick-and-place task. They found a slight advantage for the HTC Wive.

Jay et al. [16] studied delay in haptic and visual feedback in collaborative virtual environments. They found the latency in visual feedback had strong influence on the haptic task performance. They studied the effect on task requiring continuous haptic and visual exchange between participants to acquire a target.

Jay and Hubbold [17] investigated if visual and/or haptic delay influenced task performance in reciprocal tapping tasks. They found that the haptic delay had low influence, but the visual delay and combined delay had considerable impact.

Knörlien et al. [18] studied the influence of visual and haptic delay on stiffness perception in AR. They found that haptic delay decreased stiffness perception whereas visual delay increased it.

Our work is unique in the sense that the simulator provides the experience of same real-world scenario as the simulator is

digital clone of actual product in the market. In addition to this, the study includes participants from both academia and industry.

Method

Two larger formal subjective studies have been performed: one with the VR-system as it is and one where we have added controlled delay to the screen update and to the joystick signals. The former has been named the baseline experiment and the latter the delay experiment,

Common procedures for the formal tests

Test subjects were invited to perform a log-loading task in the VR simulator. They were asked the read the instructions, which explained the task to perform. It also gave a description on how to operate the crane in the simulator. As the test subjects were not required to have any previous experience in real truck crane operation, the instructions on how to move the crane with the two joysticks, see Figure 2, were somewhat lengthy, but all participants did understand this quickly when trying in the training session.

In the instructions the following was pointed out:

"For some people, an immersive simulator may give some discomfort or nausea. If you want to stop and not finish the test you can do it at any time without giving a reason. All the data that are gathered during the test will be treated and analysed strictly anonymously. We do not keep record on who is participating in the test that can be connected to the data."

We did not collect a written informed consent form, but the test leader verbally made certain that this part was clearly understood.

Figure 2: The two joysticks for operating the crane in the VR simulator and the HMD of the brand Oculus Rift

The test subjects were then asked to fill in a questionnaire with a few general questions about their experience in operating truck cranes and in using VR.

A Simulator Sickness Questionnaire (SSQ) [19, 20] was administered. This questionnaire containing 16 symptoms that were identified by Kennedy et al (1993) as relevant for indicating simulator sickness. The symptoms are:

1. General Discomfort
2. Fatigue
3. Headache
4. Eye Strain
5. Difficulty Focusing
6. Increased Salivation
7. Sweating
8. Nausea
9. Difficulty Concentrating
10. Fullness of Head
11. Blurred Vision
12. Dizzy (Eyes Open)
13. Dizzy (Eyes Closed)
14. Vertigo
15. Stomach Awareness
16. Burping

For each of the symptoms four different level of response is possible i.e. None, Slight, Moderate and Severe. The test subjects were asked to put on the HMD and adjust the sharpness of the image if necessary. Then the training session started. The task for the training session was to load two logs onto the truck. If something was still unclear, the test subjects were free to ask, and the test leader tried to answer them to make sure that the task and operation of the crane were completely clear to the test subjects.

After the main tests, a questionnaire with a couple of question about their impression of the system was filled in and then the SSQ once more.

Apparatus

The simulator is designed for training new customers and performing user experience studies related to the actual product. The simulator includes VR goggles (Oculus Rift) which provide stitched stereo cameras views, joysticks for controlling the crane and a simulation environment of lifting logs onto the truck. The computer used is a VR-ready ASUS ROG Strix GL702VM GTX 1060 Core i7 16GB 256GB SSD 17.3". The simulation software environment was built in Unity 2017.3. The input signals from the Joysticks are converted by a special purpose interface card to give game pad signals over USB. It was estimated from the simulator software developer that the delays in the baseline system were about 25 ms in the screen update from the movement of the head to rendering and about 80 ms from movement of Joysticks to visual feedback on the screen.

Baseline experiment

For the baseline experiment, although no specific demands on any specific visual ability were specified before the study, the test subjects' vision was investigated and noted down, by performing a Snellen visual acuity test, 14-chart Ishihara color blind test and a Randot stereo acuity test. The dominant eye was also investigated and noted down.

The main task consisted of loading two piles of logs with 16 of logs onto the truck. When one pile was completed the test subjects had a short break, the test leader noted down the task completion time and restarted the program. This task took about 15 minutes for one pile of logs.

After the main task was completed, the experience was investigated by letting the test subject indicate it on five rating scales shown in Figure 3. They have been constructed so that the distance between the different levels are equal distance for fulfilling interval scale properties.

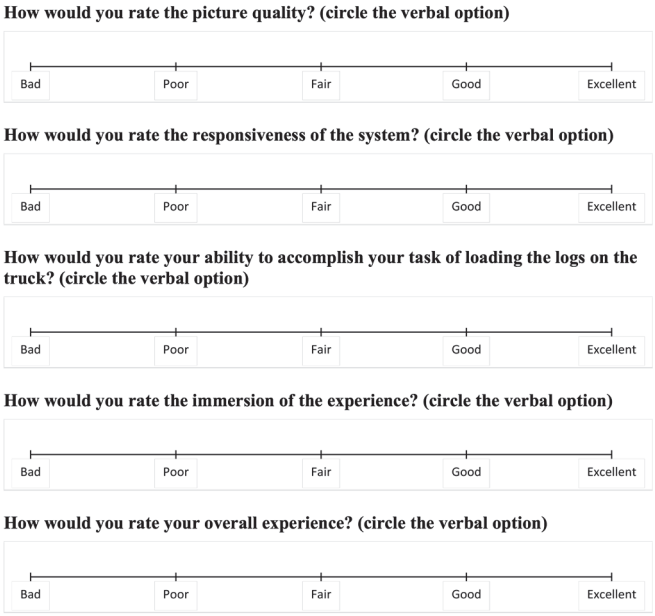

Figure 3: the rating scales used to investigate the Quality of Experience (QoE).

Delay experiment

For the delay experiment, we simplified the procedure for the visual abilities by letting the test subjects self-report their visual status.

The training session was conducted in the same way as before. It had the added purpose of giving the test subjects a sense of the baseline-delay case and this was pointed out in the instructions. The main task was to load logs onto the truck for about 20 min. The delay of the screen update and the Joysticks were adjusted every 2 min and the test subject was asked to give his or her quality ratings verbally after about 1.5 min (in practice it turned out that more time was needed to give the response so almost the whole second minute was used for that). The scales used were the same as for the baseline test, see Figure 3, except that we added one about the experienced comfort. They were also shown as sentences as below:

- How would you rate the picture quality?
- How would you rate the responsiveness of the system?
- How would you rate your ability to accomplish your task of loading the logs on the truck?
- How would you rate your comfort (as in opposite to discomfort)?
- How would you rate the immersion of the experience?
- How would you rate your overall experience?

A graphical representation of the scale was shown after these sentences, see Figure 4, in the instructions, in order to give the test subjects a mental picture of the scale.

Figure 4: Scale used in the Delay test

When the test subject was giving their ratings verbally they gave the response with the category labels: Bad, Poor, Fair, Good and Excellent.

Ten delay conditions were used (nine with added delay and one baseline-delay). These were:

- Reference condition: baseline-delay (25 ms for Display and 80 ms for Joystick)
- Display delay (ms): 5, 10, 20 and 30
- Joystick delay (ms): 10, 20, 50, 100 and 200

The order was randomized per test subject.

Analysis

Scale analysis

The scale responses were given numerical values when analyzed using the following: Bad = 1, Poor = 2, Fair = 3, Good = 4 and Excellent = 5. The instructions showed a graphical representation of the scales with equal distances between the categories. It was also pointed out in writing in the instructions. We have, therefore, assumed that we can analyze the scales as interval scales. The mean opinion scores (MOS) were calculated from the scale responses of the test subjects.

For the baseline experiment no statistical test were performed on the MOS. For the delay study comparisons and statistical test between all involved conditions were performed for each scale and delay type separately. For the Display delay we have 5x4/2 = 10 comparisons and for the Joystick delay 6x5/2 =15 comparisons.

We adopted the Bonferroni method [21] for compensating for multiple comparisons. In this method the considered significance level (α) is divided by the number of comparisons (n) so that the significance level for each comparison will be α/n. For the Display delay with 10 comparisons and $\alpha = 0.05$, we used $p \leq 0.005$ as the per comparison significance level. For the Joystick delay with 15 comparisons and $\alpha = 0.05$, we used $p \leq 0.0033$ as the significance level. The statistical test performed was with dependent T-test for paired samples. Furthermore, the hypothesis was that the delay would degrade the quality, so we have used a the one-tailed version of the T-test.

SSQ analysis

The questionnaire answers were translated into a number in our case by None = 0, Slight = 1, Moderate = 2, Severe = 3 for allowing parametric statistical analysis

Kennedy et al. (1993) [7] suggested a statistical analysis for the SSQ by grouping the different symptoms into three groups: Nausea (N), Oculomotor (O) and Disorientation (D). They also calculated a total score (TS). The Nausea symptom group contained the symptoms nausea, stomach awareness, increased salivation and burping. The Oculomotor grouped eyestrain, difficulty focusing, blurred vision, and headache. The symptom group Disorientation included the symptoms dizziness and vertigo. They are not completely disjoint since a few of the variables are used when calculating the scores in more than one group e.g. nausea and difficulty concentrating. In Table 1 it is indicated which of the symptoms that are grouped together. The calculation is done by adding together the values with a 1 in Table 1 and then

multiply that sum with factors at the bottom of the table, using the conversion between severity and numbers described above.

Table 1: SSQ score calculations as described in Kennedy et al. (1993)[19]

		Weight		
	SSQ Symptoms	N	O	D
1	General Discomfort	1	1	
2	Fatigue		1	
3	Headache		1	
4	Eye Strain		1	
5	Difficulty Focusing		1	1
6	Increased Salivation	1		
7	Sweating	1		
8	Nausea	1		1
9	Difficulty Concentrating	1	1	
10	Fullness of Head			1
11	Blurred Vision		1	1
12	Dizzy (Eyes Open)			1
13	Dizzy (Eyes Closed)			1
14	Vertigo			1
15	Stomach Awareness	1		
16	Burping	1		
	Total	[1]	[2]	[3]

$$N = [1] \times 9.54$$

$$O = [2] \times 7.58$$

$$D = [3] \times 13.92$$

$$TS = ([1] + [2] + [3]) \times 3.74$$

After the symptom scores have been calculated. The mean over the test subjects were calculated for the SSQ administered before the experiment and for the SSQ administered after the experiment.

The number of interesting comparisons performed were between each symptom group before and after, which is in total four comparisons. This gives with $\alpha = 0.05$ $p \leq 0.0125$ as the significance level. The statistical test performed was also here performed with a one-tailed dependent T-test for paired samples.

Results

Baseline experiment

The Baseline experiment was conducted at RISE Acreo AB, in a lab room for subjective experiments. No particular control of environment was done for this experiment, other than to keep it quiet from disturbing noises and at a comfortable temperature. Eighteen test subjects internally recruited from RISE[1] in Kista in Sweden, completed the test, 12 males and 6 females, with a mean age of 40 and youngest participants was 23 and the oldest 58. All but two test persons had normal or corrected to normal visual acuity. The two had a bit lower visual acuity, but could still perform the task without any problems. Two other persons were either color blind or a lower ability to see colors. The 3D acuity varied between 4 and 10, with a mean about 8.

[1] RISE (www.ri.se) is the mother organization of RISE Acreo AB

The mean and 95% confidence intervals of the ratings of the different scales used can be seen in Figure 5, see also Figure 3.

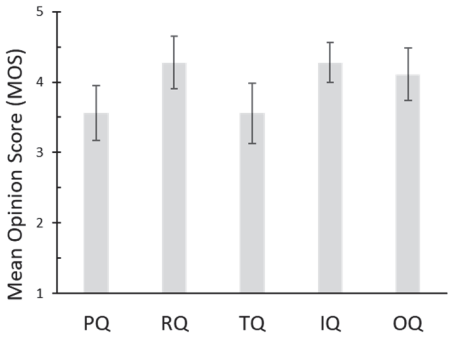

Figure 5: The Mean Opinion Scores (MOS) for the baseline experiment. From the left along the x-axis the Picture Quality (PQ), Responsiveness Quality (RQ), Task accomplishment Quality (TQ), Immersive Quality (IQ) and Overall Quality (OQ) are shown. The error bars indicate 95% confidence intervals.

The Picture Quality (PQ) was experienced as between Fair and Good. For the Responsiveness Quality (RQ) the scores were higher and the mean was just above Good. The Task accomplishment Quality (TQ) was also rated between Fair and Good. The Immersive Quality (IQ) and Overall Quality (OQ) were experienced as high, which means higher than Good.

The task completion time was 26.5 min with a standard deviation of 8.7 min.

The SSQ showed only a minor increase in the symptom strength, see Figure 6. However, the statistical test show significant increase for Disorientation $p = 0.004 < 0.01$. The other symptom groups were not significant with Nausea having $p = 0.03$, Occulomotor having $p = 0.17$ and the Total score $p = 0.02$. Most test persons reported only slight symptoms if any, and only one test person reported a moderate symptom. One interesting case of a very sensitive person was encountered. The person in question did, just after 2 min, report Severe discomfort, Nausea, Vertigo and Stomach awareness, as well as Moderate Sweating and Dizziness with eyes open. This person was not participating in the actual test, but tested the simulator in a demo session. It seems that there is a small fraction of very sensitive persons, but the majority have no major problems with this simulator.

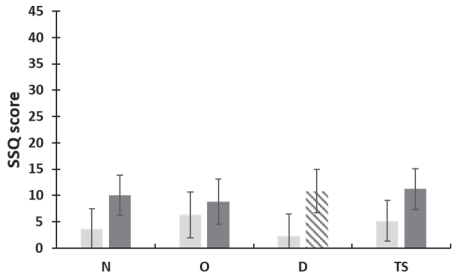

Figure 6: Simulator Sickness Questionnaire (SSQ) scores for the baseline experiment, where the left (light grey) bars represent the symptom levels before the experiment and the right (dark grey and striped indicating statistically significant difference) bars the symptom levels after the experiment. The different symptom groups along the x-axis are: Nausea (N), Occulomotor (O), Disorientation (D) and the Total Score (TS). The error bars indicate 99% (left) confidence intervals.

Delay experiment

The Delay experiment was conducted at Mid Sweden University, in a regular office room. No particular control of the environment was done for this experiment, other than keep it quiet from disturbing noises and at temperature that was comfortable. Thirty-five test subjects participated in the test, 26 males and 9 females, with a mean age of 39 and youngest participants was 23 and the oldest 61. They were recruited from the University and were a mixture of students and staff. The visual status of the test subjects was self-reported. There were a mixture of test subjects having correction and no-correction. No problems with performing the task was reported due to bad vision.

Ten test subjects stopped the test and did not complete all test conditions. The reason to stop was discomfort and nausea. In most cases, this was related to the experience of higher added Display delay conditions just before stopping i.e. added Display delay ≥ 20 ms with baseline delay ≥ 45 ms. The test leader was present during the whole test and could monitor and could also give feedback to test subject to continue or not if they felt discomfort or nausea. The recommendation was in most cases to stop. The ratings given up to the point of stopping have been included in the analysis and the ratings not given have been treated in the analysis as missing data. In all cases the SSQ were filled in for these test subjects, so these scores have been included in the analysis.

The results from the rating scales are shown in Figure 7 to Figure 12. To the left in the figures the MOS for different Display delays (DD) are drawn and to the right the MOS for different Joystick delays (JD). The total delays are given in the graphs, that is baseline delay plus added delay. For DD it is 25 ms + (5 ms, 10 ms, 20 ms, 30) = 30 ms, 35 ms, 45 ms and 55 ms. For JD it is 80 ms + (10 ms, 20 ms, 50 ms, 100 ms, 200 ms) = 90 ms, 100 ms, 130 ms, 180 ms and 280 ms. The error bars indicate 99% (left) and 99.3% (right) confidence intervals respectively. The MOS is shown pairwise where the bar to the left in each pair (light grey) is the MOS for the baseline-delay case and the bar to the right (dark grey or striped) are the MOS with delay. The MOS and error bars vary between the different pairs even for the baseline-delay case. The figures illustrate the MOS and confidence intervals obtained in the pair wise statistical tests, where the number of data pairs could be slightly different between different pairwise comparisons. The reason for this is that the test subjects that stopped and did not complete their experiment only provided partial data and since these have been included in the analysis the number of data pair vary between the comparisons.

In Figure 7, the MOS of the Picture Quality is shown. There is a trend for lower PQ at higher DDs, there is no clear trend for the JDs. A little bit surprisingly, 20 ms added Display delay (45 ms) was rated worse than 30 ms (55 ms) added delay. However, no difference could be determined significant.

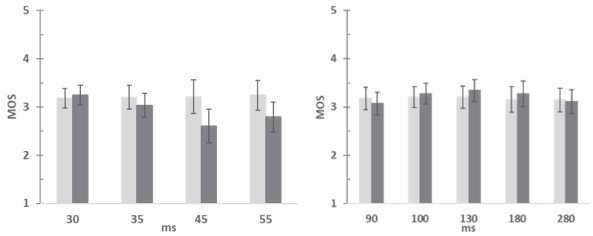

Figure 7: The Mean Opinion Scores (MOS) for Picture Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). Light grey bars are showing the baseline-delay MOS and dark grey the delay MOS. The error bars indicate 99% (left) and 99.3% (right) confidence intervals respectively.

In Figure 8, the MOS of the Responsiveness Quality is shown. There is trend for lower RQ at higher delays, but the differences are not significant.

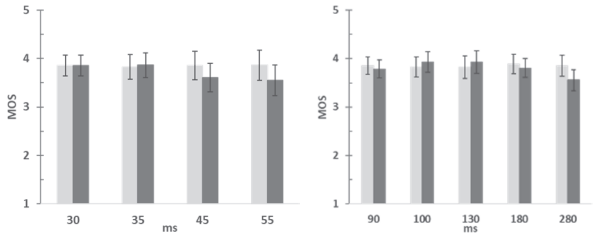

Figure 8: The Mean Opinion Scores (MOS) for Responsiveness Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). Light grey bars are showing the baseline-delay MOS and dark grey the delay MOS. The error bars indicate 99% (left) and 99.3% (right) confidence intervals respectively.

In Figure 9 the MOS of the Task accomplishment Quality is shown. No clear trend can be noticed.

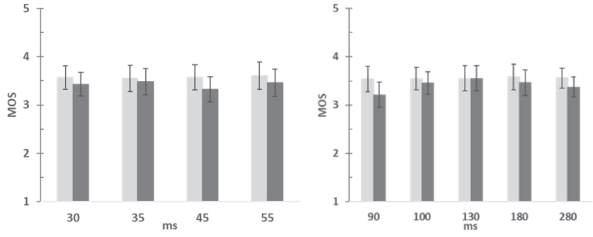

Figure 9: The Mean Opinion Scores (MOS) for Task accomplishment Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). Light grey bars are showing the baseline-delay MOS and dark grey the delay MOS. The error bars indicate 99% (left) and 99.3% (right) confidence intervals respectively.

In Figure 10 the MOS of the Comfort Quality is shown. The comfort is reduced by longer delay and this trend is clearer for the Display delay. The 30 ms added Display delay (55 ms) is significantly lower ($p = 0.0019 < 0.005$), than the comfort quality for baseline-delay.

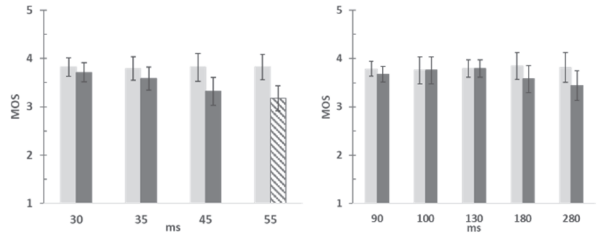

Figure 10: The Mean Opinion Scores (MOS) for Comfort Quality for different Display delay s (left) and for different Joystick delays (right) in milliseconds (ms). The 30 ms added Display delay (55 ms) was statistically significant and is highlighted with a striped pattern. Light grey bars are showing the baseline-delay MOS and dark grey the delay MOS. The error bars indicate 99% (left) and 99.3% (right) confidence intervals respectively.

In Figure 11 the MOS of the Immersion Quality is shown. There is trend for lower Immersion Quality at higher delays. 30 ms added delay IQ (55 ms) is significant ($p = 0.001 > 0.005$) compared to the baseline-delay case.

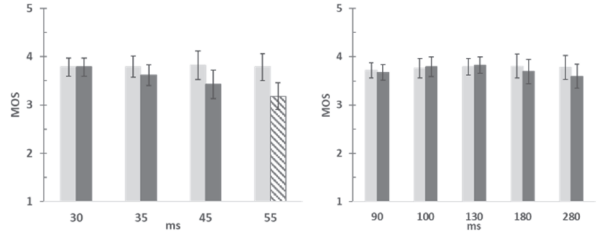

Figure 11: The Mean Opinion Scores (MOS) for Immersion Quality for different Display delay s (left) and for different Joystick delays (right) in milliseconds (ms). The 30 ms added Display delay (55 ms) was statistically significant and is highlighted with a striped pattern. Light grey bars are showing the baseline-delay MOS and dark grey the delay MOS. The error bars indicate 99% (left) and 99.3% (right) confidence intervals respectively.

In Figure 12 the MOS of the Overall Quality is shown. The OQ has similar trend as the IQ but not so clear. No significance was found.

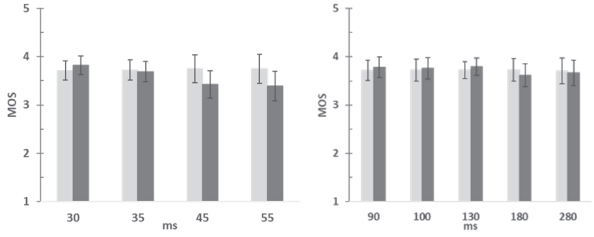

Figure 12: The Mean Opinion Scores (MOS) for Overall Quality for different Display delay s (left) and for different Joystick delays (right) in milliseconds (ms). Light grey bars are showing the baseline-delay MOS and dark grey the delay MOS. The error bars indicate 99% (left) and 99.3% (right) confidence intervals respectively.

The SSQ analysis for the delay revealed large increase in the symptom levels (Figure 13), all of which were statistically significant i.e. < 0.0125; where Nausea had $p = 0.00005$, Oculomotor $p = 0.007$, Disorientation ($p = 0.00008$) and the Total Score $p = 0.0002$). However, only 2 test subjects report symptoms on Severe level. In this analysis all test subjects were included, even those not finishing the main session.

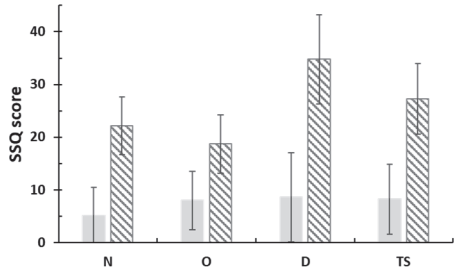

Figure 13: Simulator Sickness Questionnaire (SSQ) scores for the delay experiment, where the left (light grey) bars represent the symptom levels before the experiment and the right (dark grey and striped indicating statistically significant difference) bars the symptom levels after the experiment. The different symptom groups along the x-axis are: Nausea (N), Oculomotor (O), Disorientation (D) and the Total Score (TS). The error bars indicate 99% (left) confidence intervals.

Discussion

Baseline experiment

The scale data indicates that the test subjects are not completely satisfied with Picture Quality (MOS = 3.6 i.e. between Fair and Good).

The responsiveness is no problem and should not be since the simulation is running on a powerful enough PC with a RQ-score exceeding Good (MOS = 4.3).

For the Task accomplishment Quality, which was between Fair and Good (MOS = 3.6) i.e. most people indicating a score somewhere in the middle. Our interpretation is that the test subjects did not have a strong opinion due to minimal experience in how a real system is working (indicated in the pre-questionnaire).

Both the Immersive Quality (MOS = 4.3) and the Overall Quality (MOS = 4.1) were rated high i.e. exceeding Good.

The SSQ indicates very a little effect, although the disorientation symptoms group was significantly higher after the test as compared to be before. A small fraction of people can be very sensitive though.

Delay experiment

In the delay experiment we find some impact on lower quality for higher delay, but the effect is relatively small and we only find significant effects on the highest level on added Display delay (30 ms.) for Comfort Quality and Immersion Quality. One explanation for this seemingly small effect is that the scale analysis includes very few data samples from test subjects that did not finish the test. A reasonable assumption is that these test subject would have rated the quality lower.

Another explanation is that the task was not sensitive enough to the delay in the range in current study. Earlier studies have shown that impact of delay on task performance is very task dependent, see e.g. [16, 17]. Furthermore, test subject may not always clearly identify the delay as the source of the problem, as has been shown in telemeeting applications [22]. It can be noticed in the ratings from the test subject that several inversions exist i.e. that a test subject has rated lower quality of case with shorter delay compared to the case with longer delay.

The SSQ show a significant increase of symptoms. This is most likely connected to the Display delay, as it was seen by analyzing when test subjects were stopping the experiment it was connected to the highest added Display delay and 30 ms added Display delay also had a statistically significant lower comfort quality. The SSQ score included all participants even those that stopped, but the CQ was with a few exceptions based on the test subjects completing the test.

There was very little impact by the added Joystick delay. We can see tendencies to lower MOS on longer delays. However, no significant effects were found for the scales and as we attributed the significant effects on symptoms of SSQ to the Display delay. The Joystick delay had less impact, although we cannot analyze out the relative contributions of the two different delays.

It is known from the operation of the real crane system that the crane operators are normally very good in compensating for a bit of delay in the crane controls, which is the Joysticks in this study. It is therefore reasonable to assume that also novice operators can manage also compensate for some delay when operating the crane. Furthermore, the baseline delay as fairly long 80 ms, so the shorter added Joystick delays are relatively small and could get unnoticed just because of that

The actual undisturbed log loading time became shorter than we anticipated when planning and testing the experiment, as most test subjects needed almost 1 min to answer their ratings, which is longer than for instance when giving scores on e.g. paper or a computer interface. It may have contributed to giving less influence on the experienced delay. However, one minute is still enough time to get quite a good understanding of the environment and the test subjects were fully immersed during the rating period and continuing performing their task, so we believe it had a minor influence, but intend to investigate this further.

Conclusions

The baseline study show that most people are more or less happy with the VR-system and that it does not have strong effect on any symptoms as listed in the SSQ. There are some room for improvement since all scales were not above Good (> 4). For instance, the Picture Quality had only a MOS of 3.6.

In the delay study we found significant effects on Comfort Quality and Immersion Quality for higher Display delay (30 ms), but very small impact of Joystick delay. Furthermore, the Display delay had strong influence on the symptoms in the SSQ, as well as causing test subjects to decide not to continue to the end with the experiments, and this was also found to be connected to the longer added Display delays (≥ 20 ms).

Acknowledgement

The economic support from the Knowledge Foundation (grant nr 20160194) is hereby gratefully acknowledged. The work by Sidra Muneer in being the test leader for the Delay experiment is also gratefully acknowledged.

References

[1]. Okura, F., M. Kanbara, and N. Yokoya. *Augmented Telepresence Using Autopilot Airship and Omni-directional Camera*. in *IEEE International Symposium on Mixed and Augmented Reality 2010*. 2010. Seoul, Korea: IEEE Xplore. p. 259-260.

[2]. Saxena, V.V., T. Feldt, and M. Goel. *Augmented Telepresence as a Tool for Immersive Simulated Dancing in Experience and Learning*. in *The India HCI 2014 Conference on Human Computer Interaction* 2014. ACM New York, NY, USA. p. 86-89.

[3]. Le Callet, P., S. Möller, and A. Perkis. (2012). *Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)* (Version 1.2 (http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf)): Lausanne, Switzerland.

[4]. Möller, S. and A. Raake, *Quality of Experience - Advanced Concepts, Applications and Methods*. T-Labs Series in Telecommunication Services. 2014, Switzerland: Springer International Publishing.

[5]. ITU-R. (2012). *Methodology for the subjective assessment of the quality of television pictures* (ITU-R Rec. BT.500-13). International Telecommunication Union, Radiocommunication Sector.

[6]. ITU-T. (1999). *Subjective video quality assessment methods for multimedia applications* (ITU-T Rec. P.910). International Telecommunication Union, Telecommunication standardization sector.

[7]. De Moor, K., M. Fiedler, P. Reichl, and M. Varela. (2015). *Quality of Experience: From Assessment to Application (Dagstuhl Seminar 15022)* (DOI: 10.4230/DagRep.5.1.57 (http://drops.dagstuhl.de/opus/volltexte/2015/5036/)). DROPS (Dagstuhl Online Publication Service).

[8]. ITU-T. (2014). *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment* (ITU-T Rec. P.913). International Telecommunication Union, Telecommunication standardization sector.

[9]. ITU-T. (2016). *Display requirements for 3D video quality assesment* (ITU-T Rec. P.914). International Telecommunication Union.

[10]. ITU-T. (2016). *Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video* (ITU-T Rec. P.916). International Telecommunication Union.

[11]. ITU-T. (2016). *Subjective assessment methods for 3D video quality* (ITU-T Rec. P.915). International Telecommunication Union.

[12]. Puig, J., A. Perkis, F. Lindseth, and T. Ebrahimi. *Towards an efficient methodology for evaluation of quality of Experience in augmented reality*. in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*. 2012. Melbourne, Australia: IEEE Xplore. p. 188-193.

[13]. Debattista, K., T. Bashford-Rogers, C. Harvey, B. Waterfield, and A. Chalmers, *Subjective Evaluation of High-Fidelity Virtual Environments for Driving Simulations.* IEEE Transactions on Human-Machine Systems, 2018. **48**(1): p. 30-40.

[14]. Strazdins, G., B.S. Pedersen, H. Zhang, and P. Major. *Virtual reality using gesture recognition for deck operation training*. in *OCEANS 2017 - Aberdeen*. 2017.

[15]. Suznjevic, M., M. Mandurov, and M. Matijasevic. *Performance and QoE assessment of HTC Vive and Oculus Rift for pick-and-place tasks in VR*. in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. 2017.

[16]. Jay, C., M. Glencross, and R. Hubbold, *Modeling the effects of delayed haptic and visual feedback in a collaborative virtual environment.* ACM Trans. Comput.-Hum. Interact., 2007. **14**(2): p. 8.

[17]. Jay, C. and R. Hubbold. *Delayed visual and haptic feedback in a reciprocal tapping task*. in *First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics Conference*. 2005.

[18]. Knorlein, B., M.D. Luca, and M. Harders. *Influence of visual and haptic delays on stiffness perception in augmented reality*. in *2009 8th IEEE International Symposium on Mixed and Augmented Reality*. 2009.

[19]. Kennedy, R.S., N.E. Lane, K.S. Berbaum, and M.G. Lilienthal, *Simulator Sickness Questionnaire: An Enhanced Method of Quantifying Simulator Sickness.* The International Journal of Aviation Psychology, 1993. **3**(3): p. 203-220.

[20]. Brunnström, K., K. Wang, S. Tavakoli, and B. Andrén, *Symptoms analysis of 3D TV viewing based on Simulator Sickness Questionnaires.* Quality and User Experience, 2017. **2**(1): p. 1-15.

[21]. Maxwell, S.E. and H.D. Delaney, *Designing experiments and analyzing data : a model comparison perspective*. 2nd ed. 2003, Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc.

[22]. ITU-T. (2016). *Telemeeting assessment - Effect of delays on telemeeting quality* (ITU-T Rec. P.1305). International Telecommunication Union, Telecommunication standardization sector.

Author Biography

Kjell Brunnström, Ph.D., is a Senior Scientist at Acreo Swedish ICT AB and Adjunct Professor at Mid Sweden University. He is an expert in image processing, computer vision, image and video quality assessment having worked in the area for more than 25 years. Currently, he is leading standardization activities for video quality measurements as Co-chair of the Video Quality Experts Group (VQEG). His current research interests are in Quality of Experience for visual media in particular video quality assessment both for 2D and 3D, as well as display quality related to the TCO requirements.

Mårten Sjöström received the M.Sc. degree in electrical engineering and applied physics from Linköping University, Sweden, in 1992, the Licentiate of Technology degree in signal processing from KTH, Stockholm, Sweden, in 1998, and the Ph.D. degree in modeling of nonlinear systems from EPFL, Lausanne, Switzerland, in 2001. He was an Electrical Engineer with ABB, Sweden, from 1993 to 1994, was a fellow with CERN from 1994 to 1996. He joined Mid Sweden University in 2001, and was appointed an Associate Professor and a Full Professor in Signal Processing in 2008 and 2013, respectively. He has been the Head of the Computer and System Sciences with Mid Sweden University since 2013. He founded the Realistic 3-D Research Group in 2007. His current research interests are within multidimensional signal processing and imaging, as well as system modeling and identification.

Muhammad Imran, PhD., is a senior machine vision engineer at HIAB AB and adjunct assistant professor at STC research center, Mid Sweden University, Sweden. His research interests include design, development, and implementation methodologies for real-time vision systems in the context of machine vision and surveillance applications.

Magnus Pettersson is a senior manager at HIAB with responsibilities for new technologies, processes and integrated management system. His experience in the related hydraulics industry spans over 30 years and has worked at different positions ranging from design engineer to senior manager. His current research interests include leveraging new technologies; vision systems and IoT for improving existing on-road load handling products and offering new digital solutions in this area.

Mathias Johanson, Ph.D., is R&D manager at Alkit Communications AB and an expert on video-mediated communication and distributed collaborative environments. His research interests also include automotive telematics and e-health systems and services.