Using Electroencephalography and Subjective Self-Assessment to Measure the Influence of Quality Variations in Cloud Gaming

Justus Beyer, Richard Varbelow, Jan-Niklas Antons, Sebastian Möller

Quality & Usability Lab, Technische Universität Berlin, Germany
justus.beyer@qu.tu-berlin.de, mail@rivar.be, jan-niklas.antons@tu-berlin.de, sebastian.moeller@telekom.de

Abstract—Subjective self-assessment is a frequently used method when it comes to Quality of Experience assessment in general. If correctly used a variety of scales can be applied to assess the quality of cloud gaming for different constructs such as experienced quality or flow. Besides self-assessment, physiological correlates are a promising method to measure the influence of cloud gaming quality on the player without the interruption introduced by the rating task. We present subjective and physiological results and lessons learned from a laboratory study in which 32 participants played a first person shooter game in a cloud gaming setup with varying levels of video quality caused by different video compression bitrates. We found that the video quality influenced the perceived quality, player experience, the valence rating, and the alpha frequency band power. It is shown that 1) subjective methods assess the quality variation and 2) physiological measures capture the influence on the player in terms of a reduced cognitive state. Taken together, test set-ups will benefit from a mixed methods approach in cloud gaming QoE testing.

Keywords—cloud gaming; quality; QoE; electroencephalography (EEG); alpha frequency band power

I. MOTIVATION AND INTRODUCTION

Driven by an unprecedented availability of devices capable of running games, playing has become a mainstream spare time activity for all groups of society. While even the simplest and cheapest smartphone’s computational capacity is sufficient for an enormous repertoire of popular games, the situation is different for stationary playing devices such as PCs or game consoles where modern games often require state-of-the-art hardware. Within these eco-systems, players are forced to upgrade their hardware in regular cycles to maintain the ability to enjoy the latest titles. During the last few years, a new game delivery paradigm has gathered momentum, which has the strong potential to break that upgrade cycle: cloud gaming. With this model games are not executed at the client, but on a powerful datacenter server. All outputs are then streamed in real-time to the player over the Internet and input commands are transmitted back vice versa. The major benefit of this approach is that the player’s device only needs to display the streamed video – a task that even older and therefore less powerful hardware is capable to do. However, since the transmitted video needs to be compressed and the user input has to be transmitted quickly to the server, the quality of the interaction with the game depends highly on the network connectivity between the client and the server.

In this paper we present a study, in which we varied one key parameter of that connection, namely the video streaming bandwidth. To conduct the study we built a cloud gaming test bed using the first-person shooter “Cube 2: Sauerbraten”1 and the open source platform GamingAnywhere [1]. The participants played two levels with two different video bitrates (low and high bit rate condition), of which one led to almost no perceptible visual degradation (high bit rate) whereas the other caused heavy blurring and blockiness (low bit rate). To measure the effects of the visual degradation we used a combination of subjective self-assessment questionnaires and physiological measures with electroencephalography (EEG).

In the following Section 2, we will outline related work and introduce basic concepts used in this paper. In Section 3 we present our methods and describe the test paradigm in detail. Section 4 contains the results from the subjective and physiological measures, which we will then discuss in Section 5. In Section 6 we give an outlook on future work.

II. RELATED WORK

Since no general framework for the assessment of online gaming experience exists, multiple different methods have been proposed to characterize the influence of the network. One approach, used e.g., in [2][3][4], is to establish performance metrics like the number of kills, deaths, points attained, etc., to gauge the success of the player under certain network conditions. Since these metrics are only implicitly related to the subjective experience of the player, as more points are not equivalent to more fun, the Mean Opinion

1 http://sauerbraten.org
score (MOS), defined by the ITU in Rec. P.800 [5], has been proposed [6] and used in numerous studies, e.g., [7][8][9]. As the weighting of different quality features for the overall quality (measured in terms of MOS) varies between players, multi-dimensional constructs of digital game experience have been developed and can be measured with purposely built questionnaires such as the Game Experience Questionnaire (GEQ) [10] (used e.g., in [11],[12]). The core part of the GEQ consists of 42 items which cover seven dimensions of player experience: Challenge, Competence, Flow, Immersion (sensory and imaginative), Tension, Positive and Negative Affect. The self-assessment manikin (SAM) [13] was developed to measure the emotional state in the three dimensions: Valence, Arousal and Control. Finally, the Karolinska Sleepiness Scale (KSS) [14] is applicable to collect data of the subjects’ wakefulness state.

As self-assessment methods like questionnaires inherently place an additional burden on test subjects and interrupt the actual game experience, researchers are working on identifying physiological correlates with experience dimensions to obtain non-interruptive and continuous measures. As an example, the electroencephalogram (EEG) has proven to be a valuable tool for research in the auditory and visual domains, as it can provide additional information about underlying processes [15],[16].

EEG measures voltage changes due to brain activity by attaching electrodes to the scalp of a participant. Since Berger developed the EEG in 1929, it has been widely used for research of physiological correlates of perceptual and attentional processes [17],[18].

EEG data can mainly be analyzed in two different ways: on the one hand, by looking at the Event-Related Potentials (ERP) which are a time-locked reaction to an external stimulus measured as a change in voltage and on the other hand, by taking a closer look at the spectrogram of spontaneous (not event-related) activity [19]. With respect to the latter, there are five different frequency ranges ascribed to specific states of the brain [19]: delta band (1–4 Hz), theta band (4–8 Hz), alpha band (8–13 Hz), beta band (13–30 Hz), and the gamma band (36–44 Hz). Activity in the delta band is mainly present during sleep, theta band activity during light sleep. Activity in the alpha band is related to relaxed wakefulness and to situations of decreased alertness. High arousal and focused attention lead to a high power in the beta and gamma bands [19].

For this study, the main focus is on variations of the alpha frequency band power, which can be used as an indicator of the player’s cognitive state. A higher power in the alpha band corresponds to a reduced cognitive state. Other frequency bands were not analyzed as the variations in cognitive state were the main aim of this study. Therefore the variation of the band power of the alpha band between 9 – 11 Hz, i.e., the center of the alpha band, due to the two video quality levels will be analyzed.

III. METHODS

To study the effects of varying visual quality on the player’s subjective experience and on physiological measures, we conducted a laboratory study in which we collected both subjective ratings using questionnaires and physiological responses using EEG.

A. Experimental setup

The experiments were conducted from 01/29/15 to 02/10/15 in a laboratory room at Technische Universität Berlin, Germany. Following ITU-T Rec. P.910 [20] and P.911 [21] the room was equipped with daylight-imitating lamps and all walls were covered with thick neutral grey sound-absorbing curtains. Test participants were seated in a non-moving chair in front of a desk upon which the test client computer, a monitor, input devices and two loudspeakers were set up (Figure 1). Equipment of g.tec medical engineering GmbH² was used to continuously record the EEG signal. The participants had to put on the g.GAMMACap² containing 16 active ring electrodes located according to the international 10-20 system (Fz, F3-4, FP1-2, Cz, C3-4, Pz, P3-4, PO3-4, Oz, O1-2) [22]. Both the grounding and the reference electrode were placed at the mastoids (bone structures behind the ear channel filled with air). The signal was amplified and digitized with the g.USBamp and recorded on a dedicated computer (Fujitsu Lifebook S761, Intel Core i7 2.7 GHz, 8GB RAM, Windows 7) using the software g.Recorder.

B. Test-bed Setup and Stimulation

The hardware foundation for the cloud gaming server was provided by a DELL PowerEdge T420 server (2x Xeon E5-2430; 12 CPU cores at 2,2GHz; 64GB RAM) placed in a server cabinet with connection to the laboratory room through a switched Gigabit Ethernet network. For the study, the server was equipped with an Nvidia Quadro FX4800 graphics card. As in a realistic usage scenario, a virtualization platform was installed on the server, Citrix XenServer v6.2. Within that virtualization we created a Windows 7 instance equipped with 4 CPU cores and 4GB RAM. The

² http://gtec.at/
physical Nvidia GPU was dedicated to the virtual machine, providing 3D OpenGL rendering capabilities to the game “Cube 2: Sauerbraten” running on the open-source cloud gaming platform GamingAnywhere (v0.7.5). Being a first-person shooter, this game is particularly fast-paced and strongly depends on the player’s ability to quickly discern visual features to recognize enemies and find his/her way through the virtual world. We created two streaming configurations for the platform. Each transmitted the H.264-compressed video with a 1280x768 resolution at 50fps and OPUS-compressed audio with a 48 kHz sampling rate. In both cases, the OPUS audio compressor was configured to output 128kbit/s. However, the video encoding bitrate differed and was set to 10 Mbit/s in the high quality (HQ) case and 1 Mbit/s in the low quality (LQ) case. Since the video compression was performed entirely in software (through FFmpeg/x264), its ‘preset’ was set to ‘ultrafast’ and the ‘tune’ parameter to ‘zerolatency’ to keep encoding latencies at bay. The provisioned CPU power was sufficient to avoid frame rate degradations due to processing bottlenecks, as the observed overall utilization of the cores stayed around 50 percent.

As a client, we used a DELL Latitude D630 laptop (Intel Core 2 Duo 2.5 GHz, 2GB RAM, Windows 7), which was connected to an external 22-inch screen (Figure 1).

Within the game, two levels (“Lost” and “Level9”) were chosen based on their game mode being a campaign and the fact that the participants could not finish the level during the sessions. A campaign in “Sauerbraten” is a separately playable level where the player has to defeat enemy monsters and progress linearly to reach the end. The participants were asked to get as far as possible which included finding buttons or computer terminals to open locked doors. The basic principle stayed the same for both levels although “Lost” had some advanced capabilities as controlling a rail with a remote control. The overall interactive delay of the cloud gaming setup was observed to be about 110ms using a high-speed (240fps) camera recording.

C. Test Procedure

Participants were recruited using a web portal for the management and acquisition of test subjects. Each experiment started with an introduction phase where the participants were informed about the test procedure, had to sign the consent form and complete the first questionnaire, collecting demographic data, gaming habits, and the emotional and wakefulness state. Subsequently, the EEG equipment was set up while the participants played a training level to get familiar with the game. After the preparation of the EEG, a baseline was recorded during which the participants were asked to fixate a spot on the curtain in front of them for two minutes, and then to keep their eyes closed for the same period of time. Two gaming sessions followed, each 20 minutes long. To minimize learning effects as far as possible, instead of repeated sessions with short levels, the participants had to play both levels until they were interrupted when the time was up. The quality levels (HQ, LQ) served as random within-subject factor and the game levels were randomized to prevent order effects. After each session, a comprehensive questionnaire had to be completed gathering data in terms of quality ratings (MOS), game experience (GEQ), and again emotional (SAM) and wakefulness state (KSS). When all questionnaires were completed, the EEG equipment was removed and the test participants were offered an opportunity to wash their hair. Finally they received financial compensation.

IV. RESULTS

Altogether 32 subjects (5 females and 27 males; mean age = 25.94 years; SD = 2.723; range = 19-31) participated in the study, of whom most (25) were students. For the analysis an ANOVA for repeated measures was calculated. As independent variable the video quality level was used. The subjective scales and the alpha frequency band power served as dependent variables. The error bar in all figures indicates a confidence interval of 95%.

A. Subjective Results

The MOS ratings (collected on a scale from 1 to 7 with a step size of 0.1, where 1 corresponds to “extremely bad” and 7 to “ideal”) for the video and audio quality show the expected difference in the subjects’ perception (Figure 2). Although the audio quality was not changed, its rating is significantly affected by the video quality ($F(1,31) = 7.926, p < .01, \eta^2 = .204$) even if not as distinct as the video quality rating itself ($F(1,31) = 210.906, p < .01, \eta^2 = .872$) respectively the combined quality of audio and video ($F(1,31) = 132.517, p < .01, \eta^2 = .810$).
For the emotional state (collected on scale from 1 to 9 with step size 1) we found a significant effect in the valence dimension of the self-assessment manikin (SAM) $F(1,31) = 18.211, p < .01, \eta^2 = .370$ - test participants felt more pleasure when playing the high quality (HQ) condition (Figure 3). There is also a tendency in the control dimension, implying a feeling of being more in control during the HQ session, albeit this effect is not significant ($F(1,31) = 3.925, p < .1, \eta^2 = .112$).

The Karolinska Sleepiness Scale (KSS) (collected on a scale from 1 to 9 with step size 0.1, where 1 corresponds to “extremely alert” and 9 to “extremely sleepy – fighting sleep”) reveals another significant effect ($F(1,31) = 5.859, p < .05, \eta^2 = .159$), namely that playing the low quality (LQ) condition leads to a slightly more tired state than the HQ session (Figure 3).

Of the 7 dimensions of the Game Experience Questionnaire (GEQ) (collected on a scale from 1 to 5 with step size 1, where 1 corresponds to “not at all” and 5 to “extremely”), 6 showed significant effects (Figure 4). When playing the HQ session the subjects felt more competent ($F(1,31) = 14.235, p < .01, \eta^2 = .315$), were more in a flow state ($F(1,31) = 5.941, p < .05, \eta^2 = .161$), experienced stronger immersion ($F(1,31) = 25.207, p < .01, \eta^2 = .448$) in the game, felt less tense ($F(1,31) = 10.722, p < .01, \eta^2 = .257$) and it affected them more positively ($F(1,31) = 24.255, p < .01, \eta^2 = .327$) than the LQ session. Only the changes to the Challenge dimension were not significant although there is a slight tendency towards being more challenged when playing at LQ.

**B. Physiological Results**

In the EEG data a significant effect for the alpha frequency band power of the electrode Oz ($F(1,27) = 4.34, p < .05, \eta^2 = .138$) was found. We considered the narrow alpha band in the interval 9-11 Hz and calculated the averaged power for all but four participants, as two of them had an overly noisy signal and the other two experienced technical issues causing reoccurring recalibrations and jammed signals. Fortunately, participants excluded from the physiological analysis are evenly distributed over the randomized quality order, so no unilateral influence could result. As can be seen in Figure 5, the power spectral density in the alpha frequency band in the range between 9 to 11 Hz is higher for the low video quality condition in comparison to the high video quality condition. All other occipital electrodes showed the same tendency but did not meet significance levels.

**V. DISCUSSION**

In this study we tested the effects of video quality variations in a cloud gaming setup using self-assessment questionnaires and physiological EEG measures. The results
show that the visual quality of the game is significantly reflected in nearly all tested measures.

As expected the MOS ratings for video quality were strongly influenced by our stimuli. However, the observed MOS levels also confirm that the chosen parameter sets were appropriate to create a high and a low quality condition. One surprising feature is the significant influence of video quality variations on audio quality ratings, even though audio quality remained unchanged throughout the study.

The SAM revealed a significant effect of the video quality on the valence of the participant’s affective state, implying that they felt less pleasure after playing the LQ condition. This finding is consistent with the ratings for the Positive and Negative Affect dimensions in the GEQ.

Besides Challenge, all other GEQ dimensions were significantly affected: Lower video quality caused less positive emotions (Positive Affect) and raised negative emotions (Negative Affect). It was less immersive and left players feeling less competent. However, the bad quality also heightened the tension and might also have caused the game to be more challenging although the latter effect was not significant. Considering the very bad quality the players had to endure in the LQ condition, the observed differences in the Player Experience dimensions are lower than expected. Apparently, even a very low level of visual quality does not completely break the underlying game principle, in that it is still tense and challenging and players could enter a state of flow.

The subjective data further showed a significant effect for the wakefulness state: The study participants felt more tired after the LQ session than after the HQ session.

This effect of tiredness was also observable in the physiological EEG data: Playing the LQ condition caused significantly higher spectral power in the alpha frequency band during the first half of that session compared to the HQ condition. While this effect was also observable in the second half of the sessions, it was less pronounced and did not reach significance level. This might imply that the longer a player played the game the less influence is exerted on the wakefulness state by the video quality. As a game is an interactive endeavor as opposed to mere passive video consumption, the player may over time adapt to the degraded visual quality, and the game’s interactive content might dominate the perception.

VI. FUTURE WORK

Obviously, there are different factors affecting the user’s cognitive state. In addition to the quality level, the time-on-task seems to be important, which leads to the question of how gaming quality and the resulting effects on the player’s state can be measured while minimizing the time-on-task effect. For this aim, the gaming sessions of the experiment should ideally be short, but it can be expected that in short sessions the player will not necessarily reach a state of flow. Further research is necessary to come up with recommendations on the ideal session length, and the organization of sessions within a quality assessment experiment.

Further work is also necessary to understand the interplay of game content and quality perception in cloud gaming scenarios. This is especially true as modern high-end titles feature more photorealistic and detailed visual output than the game used in the study. Considering the serious drop in immersion caused by the degraded visual quality seen in this study, it can be assumed that these visually more complex titles suffer from video compression even more.

ACKNOWLEDGMENT

This work was co-funded by the German BMBF, funding code 01IS12056.

REFERENCES


