Game Categorization for Deriving QoE-Driven Video Encoding Configuration Strategies for Cloud Gaming

IVAN SLIVAR, MIRKO SUZNJEVIC, and LEA SKORIN-KAPOV, Faculty of Electrical Engineering and Computing University of Zagreb

Cloud gaming has been recognized as a promising shift in the online game industry, with the aim of implementing the "on demand" service concept that has achieved market success in other areas of digital entertainment such as movies and TV shows. The concepts of cloud computing are leveraged to render the game scene as a video stream which is then delivered to players in real-time. The main advantage of this approach is the capability of delivering high-quality graphics games to any type of end user device, however at the cost of high bandwidth consumption and strict latency requirements. A key challenge faced by cloud game providers lies in configuring the video encoding parameters so as to maximize player Quality of Experience (QoE) while meeting bandwidth availability constraints. In this paper we tackle one aspect of this problem by addressing the following research question: Is it possible to improve service adaptation based on information about the characteristics of the game being streamed? To answer this question two main challenges need to be addressed: the need for different QoE-driven video encoding (re-)configuration strategies for different categories of games, and how to determine a relevant game categorization to be used for assigning appropriate configuration strategies. We investigate these problems by conducting two subjective laboratory studies with a total of 80 players and three different games. Results indicate that different strategies should likely be applied for different types of games, and show that existing game classifications are not necessarily suitable for differentiating game types in this context. We thus further analyze objective video metrics of collected game play video traces as well as player actions per minute and use this as input data for clustering of games into two clusters. Subjective results verify that different video encoding configuration strategies may be applied to games belonging to different clusters.

CCS Concepts: • **Computer systems organization** → **Cloud computing**; • **General and reference** → *Empirical studies*; • **Theory of computation** → *Unsupervised learning and clustering*;

Additional Key Words and Phrases: Cloud gaming, Quality of Experience, Game categorization, Video codec configuration strategies

ACM Reference format:
Ivan Slivar, Mirko Suznjevic, and Lea Skorin-Kapov. 2017. Game Categorization for Deriving QoE-Driven Video Encoding Configuration Strategies for Cloud Gaming. *ACM Trans. Multimedia Comput. Commun. Appl.* 0, 0, Article 0 (2017), 23 pages.
https://doi.org/0000001.0000001

This work has been supported in part by the Croatian Science Foundation under the project UIP-2014-09-5605 and the project "Information and communication technology for generic and energy-efficient communication solutions with application in e-/m-health (ICTGEN))" co-financed by the EU from the European Regional Development Fund.
Author's addresses: I. Slivar, M. Suznjevic and L. Skorin- Kapov, University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, Zagreb, Croatia.

1551-6857/2017/0-ART0
https://doi.org/0000001.0000001

1 INTRODUCTION

Cloud gaming is a type of online gaming that provides game content by delivering it from a server to a client in the form of a video stream, with game controls being sent from the client to the server [Cai et al. 2016]. Resource-heavy tasks (the execution of the game logic, rendering of the 2D/3D virtual scene, and video encoding) are performed at the powerful server, while the lightweight client simplifies client-side setup and is responsible only for executing the necessary tasks at the client (video decoding and capturing of client input). While such a game streaming paradigm significantly reduces the end client device requirements as compared to "traditional" online gaming and thus allows for the delivery of graphically-rich games to less powerful client devices, the downlink bandwidth requirements are significantly increased. Furthermore, gaming in general is a highly interactive service, thus imposing strict latency requirements. In the case of cloud gaming, meeting these requirements becomes very challenging (e.g., less than 150ms of RTT is needed for good quality of First Person Shooter games [Dick et al. 2005]), with the need to calculate game state, render the virtual scene, and encode/decode the video stream. The available time budget (i.e., 150 ms) is used by network delay, virtual world state calculation, 3D scene rendering, video encoding on the server and decoding on the client side. There is thus not enough time for the server-side video encoder to perform real-time optimization of the bandwidth size of the sent video stream amid network congestion.

With available network resources varying over time, subject to issues such as varying access network conditions or a varying number of players accessing a bottleneck link, there is a need for efficient and dynamic service adaptation strategies on the game server to meet different bandwidth availabilities (e.g., adaptation of video frame rate, bitrate, resolution). A challenge faced by cloud gaming providers is configuration of the video encoding parameters used for game streaming with respect to different network bandwidth conditions. The cloud gaming server has very limited control over network latency, apart from reducing its own sending rates to avoid filling up router queues during congestion. Hence, codec reconfiguration decisions made by the cloud gaming server (in terms of chosen target bitrate and frame rate values) are driven by measured available effective bandwidth.

Previous studies that have conducted subjective end user Quality of Experience (QoE) tests have shown that different codec configuration strategies should be considered for different game types [Hong et al. 2015; Slivar et al. 2015]. In other words, selecting the appropriate video encoding parameters for different cloud games affects the efficiency of the service adaptation in terms of the impact on QoE. While there are traditional game genre-based classifications, and certain scientific approaches in classifying games (e.g., based on camera perspective [Claypool and Claypool 2009]), missing so far is a systematic approach in differentiating between game characteristics specifically for cloud gaming. Current game genres are not defined based on a set of metrics, but more informally based on different types of game mechanics. Additionally, there are many games belonging to multiple genres which makes it hard to use existing genre classification in this approach (e.g., elements of Role Playing Games can be found in many other genres from Real Time Strategies to First Person Shooters). In this paper, we evaluate possibilities for categorizing games specifically for cloud gaming based on similar objective video and game play characteristics, as suggested in [Lee et al. 2012; Slivar et al. 2016].

To evaluate how to adapt (or reconfigure) the video encoding parameters of the game video stream in light of decreased bandwidth availability for different game categories, we conduct two controlled subjective laboratory studies involving 80 participants and three tested games. The results of the first study have been published in our previous paper [Slivar et al. 2016]. This paper extends our previous work reported in [Slivar et al. 2016] by adding the results of a second subjective

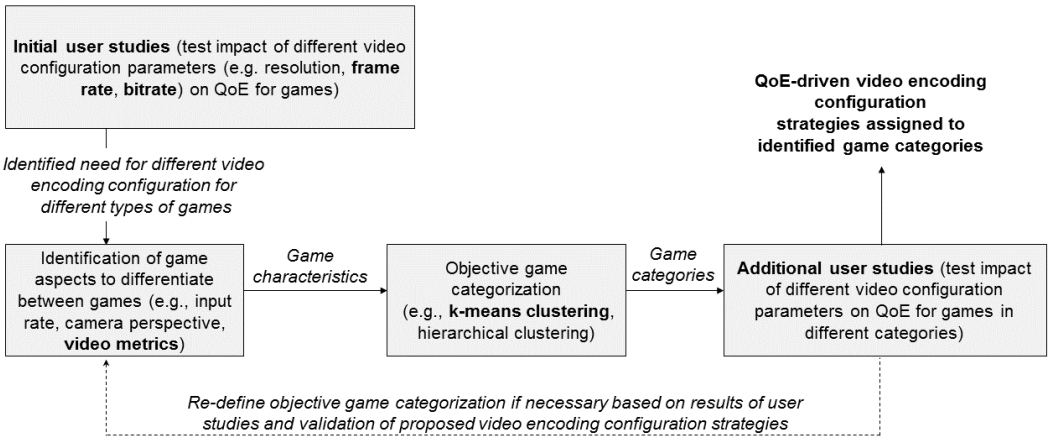


Fig. 1. Research methodology

study, and proposing a novel game categorization for cloud gaming based on objective video metrics. The results of the subjective studies are used to showcase that existing game classifications (e.g., according to game genre) are not necessarily applicable when grouping games that should have the same video encoding configuration strategy in a cloud gaming architecture. In addition to subjective scores, we report on objective video metrics aimed to characterize different games. Besides subjective studies we gathered a large number of video game play traces and collected player actions from **25 different games**. Based on a k-means analysis of obtained data, we found that **games may be grouped into 2 clusters characterized by objective video metrics**, which serves as a basis for our proposition of a **novel game categorization** that could be utilized for selecting appropriate video configuration strategies for different types of games. The overall methodology of the research presented in this paper is shown in Figure 1.

The paper is organized as follows. In Section 2 we give a comparative summary of related studies that are relevant for modeling and managing QoE in the context of cloud gaming. In Section 3 we present the methodology and the results of two subjective user studies conducted to investigate the effects of different video codec configurations on player QoE. Furthermore, motivated by the results of subjective studies, in Section 4 we report on a k-means analysis of obtained data to objectively categorize different games for the purpose of choosing an appropriate and QoE-driven video codec configuration strategy. Section 5 provides concluding remarks and directions for ongoing and future work.

2 RELATED WORK

Over the past years there have been significant research efforts in the domain of cloud gaming aimed at studying the relationships between end-user QoE and various network, service, and context factors. While many earlier studies focused on traditional online gaming have provided insight into user-level requirements in terms of factors such as perceived end-to-end latency [Claypool and Claypool 2006], cloud gaming traffic is inherently different and thus calls for new studies to determine how certain network (e.g., latency, loss) or application-level (e.g., video encoding, content) factors map to user perceived quality metrics. In Table 1 we give a detailed overview of subjective studies that have focused on measuring and modeling QoE for cloud gaming. The table contains, for each work, the information about the platform on which the tests have been

conducted, influence factors which have been tested (e.g., latency, frame rate), tested games, number of test participants in the study, measurement methodology, and identified results relevant for QoE modeling.

In terms of test platform used, numerous studies have been conducted using the GamingAnywhere platform, an open source cloud gaming system that allows researchers to perform repeatable experiments and confirm reliability of their study findings [Beyer et al. 2015; Claypool and Finkel 2014; Hong et al. 2015; Slivar et al. 2014]. Other platforms used have included Steam [Slivar et al. 2015], OnLive [Claypool and Finkel 2014; Clincy and Wilgor 2013; Lee et al. 2012; Quax et al. 2013], Ubitus [Wen and Hsiao 2014], or other experimentally set-up platforms.

With respect to tested QoE influence factors, a large number of studies have focused on the impacts of latency and/or packet loss on user perceived quality [Claypool and Finkel 2014; Clincy and Wilgor 2013; Jarschel et al. 2013; Lee et al. 2012; Liu et al. 2014; Möller et al. 2013; Quax et al. 2013; Slivar et al. 2014; Wang and Dey 2009; Wen and Hsiao 2014], while fewer studies have addressed the impact of different video encoding configurations on QoE [Beyer et al. 2015; Hong et al. 2015; Liu et al. 2014; Slivar et al. 2015; Wang and Dey 2009].

Furthermore, while certain studies are focused on developing models for estimating actual user QoE [Wang and Dey 2009], such models can be too complex (in terms of number of predictors considered) and thus not applicable in the context of application and cloud resource adaptation (i.e., when attempting to reconfigure video codec parameters on the fly) [Hong et al. 2015].

With regards to tested games, most of the studies have recognized game genre as the context factor having the most significant impact on QoE. As a result, many user studies have considered games from different game genres for conducting QoE tests [Claypool and Finkel 2014; Hong et al. 2015; Jarschel et al. 2013; Liu et al. 2014; Möller et al. 2013; Slivar et al. 2016, 2015; Wang and Dey 2009; Wen and Hsiao 2014], such as those differing in: camera perspective, graphics style and quality, game play pace, and the intensity of user interaction. As might be seen, a large number of different games have been included in the studies, wherein the differences between some of them can not be clearly identified. Commonly utilized video game genres are primarily derived based on the viewpoint used in the game and the game theme [Cai et al. 2016]. Based on the viewpoint, games are commonly categorized as first-person, third-person, and omnipresent, with each of these categories having different QoE requirements for traditional online gaming [Claypool and Claypool 2009], as well for cloud gaming [Jarschel et al. 2013; Quax et al. 2013]. In addition to these categories, there are significantly more game categories derived based on game theme, such as action, sports, fighting, racing, shooting, role-playing, and strategy games. A combination of the game viewpoint and the game theme results with numerous distinctive game genres, e.g., first-person shooter (FPS), third-person action games, and real-time strategy (RTS). For most such obtained game genres, QoE requirements for game streaming differ [Hong et al. 2015; Slivar et al. 2016, 2015], although there is an indication that for some game genres, the same adaptation policy could be utilized, as shown later in the paper. Currently, there is no systematic approach available in literature for selecting which games (or which types of games) to use when conducting QoE studies, as an appropriate game categorization (grouping together games with similar QoE requirements) at the moment does not exist.

3 USER STUDIES

3.1 Methodology

To provide a better understanding of how the video codec configuration of a game stream affects player QoE, two separate subjective studies were carried out. Both QoE studies consisted of participants taking part in a two and a half hour long gaming session that was conducted in a

Table 1. Overview of studies addressing cloud gaming QoE

Author (Year)	Platform	Tested QoE influence factors			Game genres	No. of participants; environment	QoE measurement methodology	Relevance for QoE modeling
		Network factors	Video factors	Context factors				
Hong et al (2015) [Hong et al. 2015]	GA	-	Frame rate, bitrate	Game genre	FPS, action, racing	101; crowdsourced study	7-pt. ACR scale	Proposed gaming QoE MOS model as a quadratic function of video encoding parameters
Slivar et al (2015) [Slivar et al. 2015]	Steam plat-form	-	Frame rate, bitrate	Game genre, player skill	RPG, FPS	15; controlled lab environment	5-pt. ACR scale; Overall QoE and its features, willingness to play	Modelled QoE as a linear function of video frame rate and bitrate
Claypool et al (2014) [Claypool and Finkel 2014]	OnLive & GA	Latency	-	Game genre, different type of client's device	Racing, platform	49 (OnLive), 34 (GA); controlled lab environment	7-pt. ACR scale (OnLive); 5-pt. ACR scale (GA); Game play experience	Cloud-based games are as sensitive to latency as FPS games in traditional online gaming
Slivar et al (2014) [Slivar et al. 2014]	GA	Latency, packet loss	-	Player skill	MMORPG	35; controlled lab environment	5-pt. ACR scale; Overall OoE and its degradations, willingness to play	Modelled QoE as a linear function of network delay and packet loss
Wen et al (2014) [Wen and Hsiao 2014]	Ubitus	Latency, band-width	-	Game genre, PC set-up, game special effects	FPS, action and fighting	14; controlled lab environment	5-pt. ACR scale; Video and game play smoothness, graphics quality	MOS of all measured QoE components strongly correlated with network delay
Liu et al (2014) [Liu et al. 2014]	Exper. set-up	Latency, packet loss	Frame rate, bitrate	Game genre, game content (view distance, texture detail)	FPS, RPG	18 (first study), 23 (second study); controlled lab environment	5-pt. ACR scale; CMR-MOS	Proposed a content-aware model for mobile cloud gaming
Ahmadi et al (2014) [Ahmadi et al. 2014]	-	-	-	Game genre, game content	8 different genres	20; controlled lab environment	5-pt. ACR scale	Proposed a game attention model for efficient bitrate allocation in cloud gaming
Beyer et al (2014) [Beyer et al. 2015]	GA	-	Bitrate	-	FPS	32; controlled lab environment	GEQ, EEG	Low video quality imposed by low video bitrate has significant effect on participant's satisfaction
Jarschel et al (2013) [Jarschel et al. 2013]	Exper. set-up	Latency, packet loss	-	Game genre	RPG, sports, racing	58; controlled lab environment	5-pt. ACR scale; Overall QoE, willingness to pay	Identified key influence factors for cloud gaming QoE
Quax et al (2013) [Quax et al. 2013]	OnLive	Latency	-	Game genre	RTS, platform, racing, action	8; controlled lab environment	7-pt. Likert scale & GSR; Perceived game play experience, enjoyment and frustration	Latency has similar impact on QoE for the different genres in cloud gaming as in traditional online gaming
Clincy et al (2013) [Clincy and Wilgor 2013]	OnLive	Latency, packet loss	-	-	FPS	50; controlled lab environment	5-pt. ACR scale; 8 categories of QoE used to derive QoE index;	In cloud gaming, FPS players are more sensitive to network impairments then RPG players
Möller et al (2013) [Möller et al. 2013]	Exper. set-up	Latency, packet loss, BW	-	Game genre, player skill	Action, casual	19; controlled lab environment	7-pt. ACR scale; 7 quality aspects of QoE	Complexity of activity in game scene should be considered as influencing factor on QoE
Lee et al (2012) [Lee et al. 2012]	OnLive	Latency	-	Game genre	FPS, RPG, action	15; controlled lab environment	fEMG	Proposed a game real time- strictness model based on user input rate and game dynamics
Wang et al (2009) [Wang and Dey 2009]	Exper. set-up	Latency, packet loss	Frame rate, video resolution	Game genre	Sports, MMORPG, racing	21 & 15; controlled lab environment	GMOS (Game Mean Opinion score)	Proposed a model for mobile cloud gaming user experience based on manipulated factors in the study

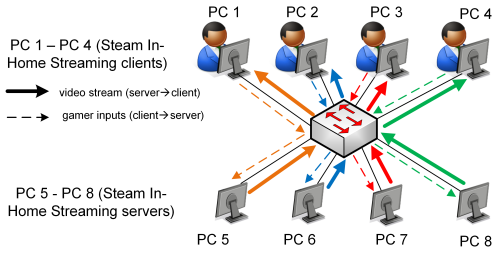


Fig. 2. Laboratory testbed

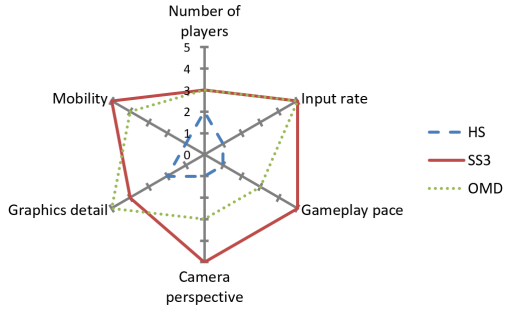


Fig. 3. Comparison of game characteristics for the tested games. Points on the different axis represent different levels as described in the paper.

laboratory environment as shown in Figure 2. Valve's Steam In-Home streaming platform was used as the cloud gaming environment [1]. The Steam client application was installed on PCs in the laboratory, thus converting PC1-PC4 (Windows 7 desktops, each with Intel 3.3 Ghz i3 processor, 4GB RAM and GIGABYTE Radeon R7 250 graphic card) to Steam In-Home Streaming clients (*cloud gaming clients*) and PC5-PC8 (Windows 8 desktops, each with Intel 3.6 Ghz i7 processor, 8GB RAM and ASUS GT740 OC graphic card) to Steam In-Home Streaming servers (*cloud gaming servers*). Each of the clients had a corresponding Steam In-Home Streaming server associated, therefore four participants were able to play simultaneously during the experiments.

With regards to the diversity of the tested games, the aim was to test games belonging to different genres so as to determine whether or not different codec configuration strategies should be applied across different types of games. Three different games were played in the user studies as follows:

- **Study 1**: *Serious Sam 3 (SS3)*, version 3.0.3.0, representing a fast paced first person shooter game, and *Hearthstone (HS)*, version 2.6.0.8834, a relatively slow paced card game,
- **Study 2**: users again played SS3,version 3.0.3.0, and as a second game they played *Orcs Must Die! Unchained (OMD)*, open beta version 1.0.18725.0, a third person hybrid action tower defense game.

We illustrate the differences between the tested games in Figure 3 and according to the following characterization dimensions (inspired by the categorization given in [Suznjevic et al. 2009]): number of players, input rate, game play pace, camera perspective, graphics detail, and mobility of avatars. The intended use of the figure is to visualize (in a straight forward manner) fundamental differences between the studied games. We note that portrayed dimensions are not necessarily orthogonal, and that not all values of these dimensions may be feasible in a game spectrum. Each dimension is divided into five levels, except for camera perspective, which is divided into three levels based on [Claypool and Claypool 2009]. The number of players is divided into five levels (from 1 to 5): single player games, two-player games, games intended for up to ten players, games intended for up to 100 players, and games for more than 100 players. In this dimension HS is placed into category 2, and SS3 along with OMD into category 3. Input rate is divided based on average action per minute rate (APM) into the following categories: <10 APM, between 10 and 20 APM, between 20 and 30 APM, between 30 and 40 APM, and 50 and more APM. In this dimension HS is placed into category 1, while other tested games are placed into category 5. The game play pace is specified based on

[1]Steam In-Home streaming, http://store.steampowered. com/streaming/

Table 2. Summary of differences between conducted user studies

Study	Year	Number of participants	Tested games
1	2015	52 (16 novice, 22 intermediate, 14 experienced)	Serious Sam 3, Hearthstone
2	2016	28 (8 novice, 9 intermediate, 11 experienced)	Serious Sam 3, Orcs Must Die! Unchained

the rate of the events in the game which require player reaction. In this dimension, HS is placed into category 1 as the pace is very low (usually players need to react to 1 or 2 events in 70 seconds). SS3 is placed into category 5 as the rate of events (i.e., attackers in the game) can be even multiple in one second, whereas OMD is plotted in the middle between these two extremes into category 3 as the pace changes over time (ranging from a slow paced placing of defences to highly paced battles with waves of enemies).

We opted to use HS and SS3 in Study 1 as they represent two ends of the spectrum on many of the defined dimensions, resulting with selection of two completely different games in every considered game play aspect. For Study 2, OMD was selected from a different game genre (according to existing game classification) as compared to HS and SS3, but is similar in some of the dimensions to SS3. All three games were played in HD-ready resolution (720p) with default graphics settings.

The participants in our subjective tests were 80 students enrolled at the University of Zagreb: in Study 1 there were 38 male and 14 female adults, aged between 21 and 26 (median age 23), while 21 male and 7 female, aged between 22 and 33 (median age 23) participated in Study 2. Prior to the experiments being conducted, the participants were instructed to fill in an online questionnaire, so as to obtain relevant information about their previous overall gaming experience and gaming experience relevant to the tested games. As a result, 16 novice, 22 intermediate skilled and 14 self-reported experienced players took part in Study 1, whereas in Study 2 the participants were on average more experienced (8 novice players, 9 intermediate and 11 experienced). Since previous studies for traditional online gaming have shown that players' group composition based on previous gaming experience has an impact on perceived QoE [Suznjevic et al. 2013], test groups were formed accordingly to investigate if this phenomenon occurs similarly in cloud gaming. The participants were organized in 13 groups (Study 1) and 7 groups (Study 2) with 4 players in each group, based on their reported gaming experience (skill). Each of the formed heterogeneous groups had one novice and one experienced player, while homogeneous groups consisted of 4 players with the same gaming skill level. One of the reasons for letting participants play together in groups was that for less experienced players, we expected it to be more interesting and enjoyable to play in groups with other colleagues. Moreover, allowing players to play in groups rather than alone may be considered more representative of a real-world scenario. However, it is important to acknowledge that controlling the social factors (communication between players and variable session length based on their performance) in this situation is more difficult and might have adverse effects on the results (note that previous studies [Suznjevic et al. 2013] showed that the quality of gaming can be rated differently when mixed player groups are used in experiments). In future studies, an option may be to use an AI or expert gamer playing as an opponent to mitigate the impact of this social influence factor. Table 2 summarizes the differences between the conducted user studies.

As stated previously, our focus in this paper is not on analysing the impact of network parameters on cloud gaming (as has been addressed in many previous studies), but rather on the

investigation of the impact of video encoding parameters on QoE, with a focus on the cloud game provider perspective. Therefore, we manipulated video frame rate and bitrate, consequently controlling/influencing image quality and smoothness of game play. Our aim was to investigate how and to what extent these parameters affect perceived QoE for different types of games, with the ultimate goal being to use this information to derive codec configuration strategies and optimize resource allocation (from a network/service provider standpoint), while at the same time preserving high QoE. For the manipulation of video frame rate, we decided to use four levels of frame rate: 25 fps, 35 fps, 45 fps and 60 fps. In the aforementioned previous studies [Hong et al. 2015; Slivar et al. 2015], the lower end of the fps spectrum was investigated, so we opted for relatively higher values of frame rate, which coincides with the expectations of average experienced gamers regarding video frame rate. As far as video bitrate is concerned, we selected three levels for the test scenarios: 3 Mbps, 5 Mbps and 10 Mbps. Both frame rate and bitrate were manipulated through Steam's developer console. The average bitrate received by the client during the experiments corresponded to the server settings, even though in some cases (e.g., low-motion video), the average bitrate was slightly lower than the settings. For encoding parameters, we note that video resolution always stays the same, while for other parameters (e.g., QP, RD) we assume that the system adapts them accordingly (e.g., lowering frame rate at the same bitrate results with better graphics quality). We note that we had to limit ourselves to a certain number of test conditions, constrained by the length of subjective testing sessions. Additional test conditions would potentially lead to overly lengthy gaming sessions and possibly player fatigue. The chosen test conditions were based on our aim to complement previous studies, in the sense that we address conditions under which the impact of different bitrate/frame rate combinations on QoE has not been well studied. Furthermore, prior to the user studies, we conducted tests to check if our testbed set-up has sufficient hardware and software capabilities necessary to support all tested games and conditions. We measured performance (frame rate) of our testbed for each tested condition and it proved sufficient for all conditions.

Considering manipulated video encoding parameters and different games, a total of **24 different test conditions** were investigated in each of the studies, with all conditions tested by each test group. During one test scenario, all players tested the same conditions (i.e., video encoding parameters). To avoid bias of manipulated video parameters, the sequence of test scenarios was randomized for each group. At the very beginning of the experiment, the participants were familiarized with the concept of cloud gaming and the Steam In-Home Streaming service. All the participants from each test group were seated in the same experimental room, with PCs located next to each other in one row (the participants could see each others screens and communicate with each other during experiments). Before tests started, the participants were given a short time to familiarize with game specific mechanics and the chosen map. It should be noted that the reference testing conditions were different between conducted studies. In Study 1, the participants played a tutorial phase on the cloud gaming servers, thus experiencing unimpaired game play by the cloud set-up. On the other hand, the participants in Study 2 initially played games under the best test conditions on cloud gaming clients (training phase).

Regardless of the previously described difference in methodology between the studies, in both studies the first 12 test scenarios involved playing one round of SS3 cooperative survival mode on a single map. During these test scenarios, participants cooperated with each other to survive longer on the map. Each of these 12 test scenarios lasted on average from 2 to 5 minutes, depending on how long players from the test group survived. After finishing each test scenario, the participants were instructed to report *overall QoE*, *perceived graphics quality* and *perceived fluidity of game play*, on a 5-pt. ACR scale (we note that such measures have also been reported in related work

[Hong et al. 2015]). Fluidity was explained as referring to the perception of the smoothness in the rendering of the virtual scene. Additionally, participants also reported their willingness to continue playing under the given test conditions for the current test scenario (yes/no). We also recorded the survival time for each player[2]. While participants were filling in questionnaires, the test administrator changed the video encoding parameters by running scripts on the player's PCs. The second half of the experiment involved playing HS or OMD, depending on the study. The group composition remained the same in the case of OMD, while for HS it changed. HS is a digital card game that consists of turn-based matches between two players. For that reason, an opponent from the group was assigned to each player by the test administrator. In the case of HS, each test scenario lasted 3 minutes, after which the participants filled in a questionnaire and continue playing the ongoing match, while the duration of test scenarios for OMD sessions lasted between 3 and 5 minutes, depending on the efficiency of a player's defences in eliminating incoming enemy waves. The entire gaming session (with a 10-minute break allotted in the middle) lasted approximately two and a half hours, depending on the group's performance during the SS3 and OMD test scenarios. We note that potential order effects may have occurred during experiments due to the experimental design (order of games).

3.2 Results

To evaluate and compare the results from the two studies, we first calculated the Standard Deviation of Opinion Scores metric (SOS), as proposed in [Hoßfeld et al. 2016]. SOS reflects the user diversity and its relation to the Mean Opinion Score (MOS). This metric is significant for this paper because we conducted two separate studies with different users and wanted to evaluate the diversity of the users involved. We calculate the SOS metric for our two studies *per test condition*, i.e., for each of the tested 24 conditions. Figure 4 (left figure) depicts the values for both studies. We also calculate the fit defined with the SOS hypothesis [Hoßfeld et al. 2011] which states that the SOS and MOS relationship can be modelled with the following equation (assuming ratings on a 5 pt. Absolute Category Rating (ACR) scale):

$$SOS(x)^2 = a(-x^2 + 6x - 5), \tag{1}$$

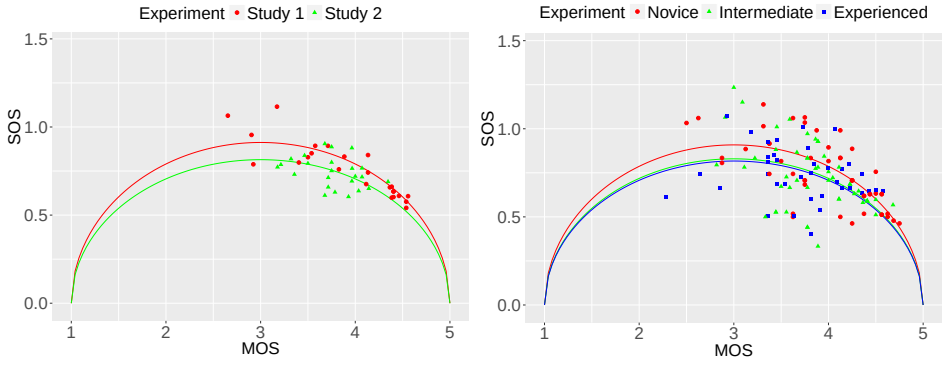
where x represents MOS.

We performed the fitting on our data and found the values of the a SOS parameter as being 0.2080 and 0.1659 for Study 1 and Study 2, respectively. This can be interpreted as meaning that the diversity of users scores was greater in our first study. We note that there were only two differences between the studies in terms of test methodology. The first difference was related to the initial training phase as previously described, whereby in Study 1 players had access to an "ideal" scenario (i.e., players played the game on a PC which was used as a server for cloud gaming testing). The second difference was in the tested games. In Study 1 we used SS3 followed by Hearthstone, while in Study 2 we used SS3 followed by OMD. When comparing the obtained a values with those reported in previous studies as summarized in [Hoßfeld et al. 2016] (e.g., 0.230 for speech QoE on a discrete scale, 0.123 for speech QoE on a continuous scale, 0.268 for Web page load times QoE, 0.099 for video QoE, and 0.266 for task related Web QoE), we can conclude that our studies have relatively low a values and that for such a variable activity as gaming this could be taken as acceptable user diversity.

Another user parameter which we evaluated was players' experience in games. We calculated the SOS parameter for each of the three defined player experience ranks: novice, intermediate, and experienced. These values were calculated per study and per test condition, leading up to 48 data points for each experience rank. We also perform fitting of the a parameter for each rank,

[2]Complete results of subjective studies with all user scores are available at www.fer.unizg.hr/qmanic/data_sets

obtaining 0.2063, 0.1720, and 0.1668 for novice, intermediate, and experienced players respectively. The data and the curves fitting each of the experience ranks are depict in Figure 4 (right figure). It can be observed that novice players have significantly more spread in giving their scores then intermediate and experienced players. This may be attributed to the fact that novice players are not that familiar with how the game should behave and possibly have greater diversity in terms of quality expectations, so their scores tend to fluctuate more from the MOS scores.



(a) Per study (b) Per experience rank

Fig. 4. Standard Deviation of Opinion Scores

As we conducted two separate, but essentially very similar user studies, we considered the possibility of combining results for SS3 to estimate a combined effect across two studies. For that purpose it was essential to check if the effects found in the individual studies are similar enough to join the data and analyze the combined effect. Therefore, we tested the homogeneity of variance of the data using Levene's test and results have shown that our SS3 group variances cannot be treated as equal ($F = 27.32, p < 0.05$). The most likely explanation for the difference between SS3 QoE ratings in both studies lies within the design of the user studies. From the methodological perspective, the only difference between conducted user studies was in the training phase: in Study 1, participants played a tutorial phase on the cloud gaming servers, thus experiencing unimpaired game play by the cloud set-up. On the other hand, the participants in Study 2 initially played games under the best test conditions on cloud gaming clients. Such a difference in terms of methodology might explain why the subjective QoE ratings for SS3 are lower in Study 1 than in Study 2. Another additional explanation for the phenomenon could be the social factor: two different groups of players participated in the studies and the social dynamics between them might affect perceived gaming quality. Consequently, scores for SS3 are analyzed separately across the two studies.

Figure 5 shows the average subjective ratings of overall QoE for SS3 and HS in Study 1 across all test conditions. First of all, it can be observed that there is a visually significant difference between overall QoE for both games: HS has on average higher scores of overall QoE for all test conditions in comparison with SS3, with the average QoE score never going below 4.0 for any given test scenario. A one-way analysis of variance (ANOVA) was used to determine whether there are any statistically significant differences between the means of two tested games. It should be noted that we consider our data as interval data and not ordinal (i.e., we consider that the intervals between points on the rating scales are equal). One-way ANOVA indeed confirmed our observation that there is statistically significant difference between QoE scores for tested games ($F = 415.26, p < 0.05$). Furthermore, average QoE scores for SS3 are significantly lower than HS QoE scores for each of

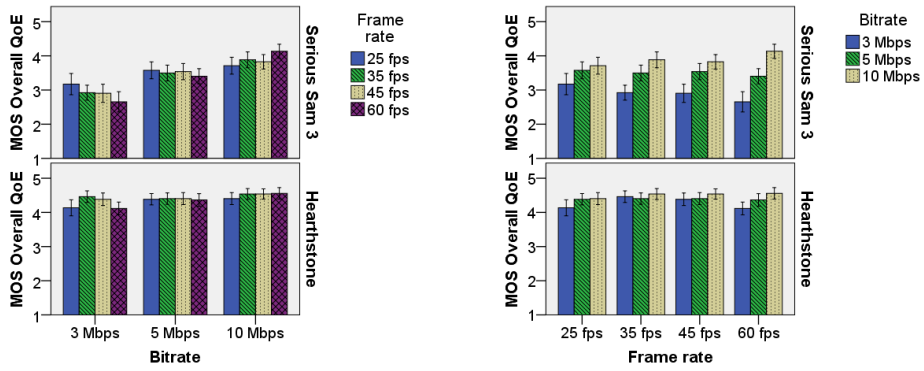


Fig. 5. Subjective ratings of overall QoE (95% CI) in Study 1

the test scenarios, with only one test scenario (the case with 60 fps and 10 Mbps) averaging more than a 4.0 score.

Moreover, it can be noticed that manipulation of video encoding parameters significantly affects perceived QoE for SS3 gaming sessions (one-way ANOVA results: $F = 13.198, p < 0.05$): when bitrate values are high enough (10 Mbps), lowering frame rate leads to degradations of QoE. SS3 is a representative fast paced first person shooter game, thus degradations of fluidity (smoothness of game play), introduced by lowering frame rate, have a higher impact when the bitrate is high enough to support transmission of high quality video. However, for low bitrate levels (3 Mbps), average scores of perceived QoE are ascending with reductions of frame rate (down to 25 fps). This can be attributed to the fact that 3 Mbps bitrate is not high enough to preserve *good enough* video quality, so even though fluidity is very important for fast paced games, the participants do not tolerate low graphics quality and thus prefer an increase in graphics quality at the expense of lowering the fluidity of game play for these scenarios. On the other hand, we observe that neither lowering video frame rate nor video bitrate had such a severe impact on perceived QoE during HS gaming sessions (average QoE score for all test scenarios is above 4.0), though the results of one-way ANOVA indicate statistically significant difference between QoE scores for different video codec configurations ($F = 2.524, p < 0.05$). We can assume that during our experiments, the manipulated frame rate and bitrate values were high enough that the participants did not perceive QoE degradations for HS.

Likewise, the average subjective ratings of overall QoE for tested games in Study 2 are shown in Figure 6. Contrary to the results from Study 1, there was no statistically significant difference between QoE scores for SS3 and OMD, as determined by one-way ANOVA ($F = 2.282, p = .131$): we can observe that QoE scores for both games have very similar values across the same test conditions. Furthermore, for both tested games, we noticed that manipulation of video encoding parameters could be utilized for achieving higher QoE levels under low network bandwidth availability. That is particularly visible for the test conditions with 3 Mbps bitrate, where it is clearly beneficial to lower the frame rate of a game stream to achieve better QoE scores. This reinforces the claim that in the case of poor network conditions, participants prefer graphics quality increase at the cost of game play fluidity.

Comparing with the results from Study 1, QoE scores for SS3 are on average higher in Study 2. For example, there are only 2 test scenarios in Study 2 that have average QoE scores lower than 3.5, while in Study 1 approximately half of the test scenarios are evaluated with such a low

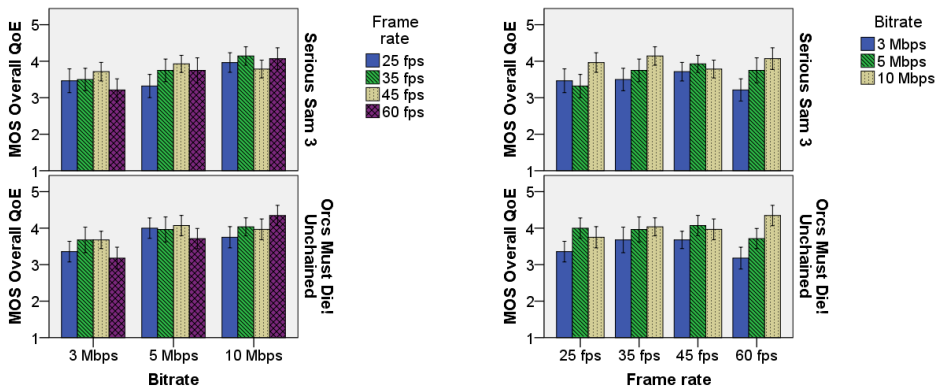
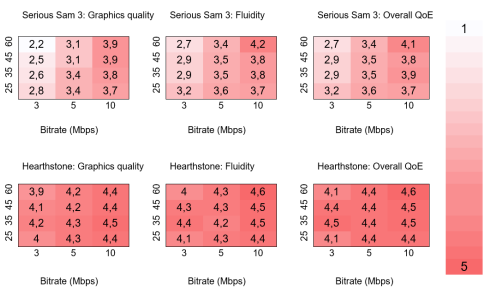


Fig. 6. Subjective ratings of overall QoE (95% CI) in Study 2

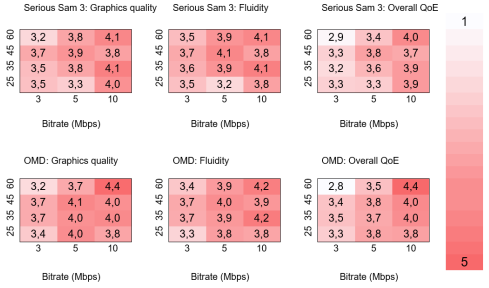
average score. These results could most likely be explained by the aforementioned difference in the reference testing conditions between the studies, resulting with higher QoE scores for SS3 in Study 2. Nevertheless, given the results from both studies, we see that *for certain games, different encoding strategies may be employed to maintain high player QoE*. This builds on and complements the results reported in [Hong et al. 2015], where different games were tested, and at different bitrate and framerate levels. More details regarding actual QoE models derived based on the results of Study 1 can be found in our previous work [Slivar et al. 2016].

Besides collecting data about overall QoE scores we collected data about user perceived fluidity and graphics quality. A heatmap overview of collected data (Figure 7) shows the mean scores for overall QoE, graphics quality and fluidity across both studies. Spearman's rank-order correlation was computed to determine the relationship between overall QoE and measured QoE dimensions. There is a very strong, positive correlation between overall QoE and fluidity ($r_s = .811, p < .001$), and overall QoE and graphics quality ($r_s = .809, p < .001$) in Study 1. Similarly, the relationship between measured scores in Study 2 is nearly identical: overall QoE and fluidity have a strong, positive correlation ($r_s = .739, p < .001$), as well overall QoE and graphics quality ($r_s = .771, p < .001$). Therefore, the strong correlation between the measured metrics indicates that players form an opinion about the test scenario and score the different dimensions based on this opinion. It can be noticed that the HS MOS score for overall QoE and its features (fluidity, graphics quality) are on average much higher and are prone to minor changes due to manipulation of video parameters in comparison with MOS scores with other tested games. This further supports the claim that the majority of players do not easily perceive QoE degradations while playing a slow paced game such as HS across a wider range of test conditions.

In addition to differences in aggregated scores, we also report on user's willingness to continue playing for each of the test scenarios as shown in Figure 8. There is a large discrepancy in the number of test scenarios where the participants were not willing to continue playing under current test conditions between tested games. In Study 1: for SS3, there were 218 occurrences (from 624 overall) when players stated they would not continue playing under the given conditions, while for HS there were only 13 cases (from 624 overall) when players stated they would not continue. It can be observed that for 3 Mbps and 60 fps, 73.1% of players were not eager to continue playing SS3, while for HS under the same test conditions only 1.9% players wanted to quit playing. Additionally, we observe that at a bitrate of 3 Mbps, a decrease in frame rate actually results in an *increase* in the percentage of players reporting they would continue playing SS3, whereas for HS the same



(a) QoE feature's ratings for Study 1



(b) QoE feature's ratings for Study 2

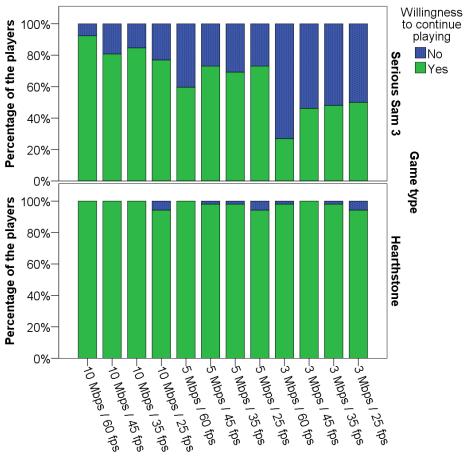
Fig. 7. Aggregated subjective ratings for each game under different video configurations

manipulation of frame rate does not result with such an increase in the percentage of players willing to continue playing.

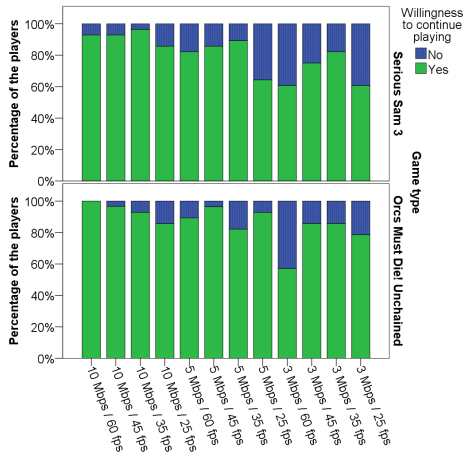
In the case of Study 2, we can notice a similar pattern of players not willing to keep playing for both tested games: the percentage of players unwilling to continue playing for SS3 changes across different test scenarios, as was also the case for OMD. At a bitrate of 10 Mbps, reducing frame rate resulted in an increase in the percentage of players not willing to continue, while at a bitrate of 3 Mbps the equivalent change of frame rate leads to a higher percentage of players that are willing to continue playing. As in Study 1, the test scenario with 3 Mbps and 60 fps was the "worst" test scenario with more than a third of players willing to quit playing for both tested games: 39.3% of players in the case of SS3 and 42.9% of players for OMD. This further confirms that the same video encoding strategy could be employed for different types of games when aiming to optimize end user QoE and reduce player abandonment of the service.

Given the length of both user studies (approximately 2.5 hrs per player), we also tested whether there was an impact on user fatigue and trends in user ratings from beginning to the end of the test session. Thus, we compared the overall user ratings for conditions tested early in the test session and those tested late in the session (note test ordering was randomized). We note that no clear differences or trends were observed.

Furthermore, we inspected the impact on QoE of the players' social context. Social context is represented by players' group composition based on previous player's gaming experience. Due to lower number of participants in Study 2, the participants could not be organized in groups that have players with the same gaming skill, therefore the impact of group composition was only investigated for Study 1. The distinction of QoE scores for SS3 between homogeneous and mixed groups is minor across all experience levels, although a slight decrease of perceived QoE can be observed when playing in mixed groups for all levels. However, group composition has a different impact on QoE for HS. While novice players report lower QoE scores in mixed groups, perceived QoE of intermediate and experienced players is slightly improved while playing in the same group composition. This can be attributed to the nature of the tested games. Whereas SS3 was played cooperatively in our study, HS is a game where two players play against each other and only one of the players wins. This sometimes results with imbalanced game sessions where novice and experienced players are paired against each other, and in these types of game situations more experienced players generally win with ease, making gaming sessions more enjoyable for winners, as also reported in [Claypool et al. 2015].



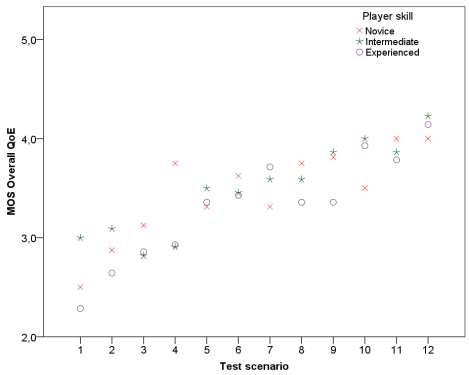
(a) Willingness to continue playing for Study 1



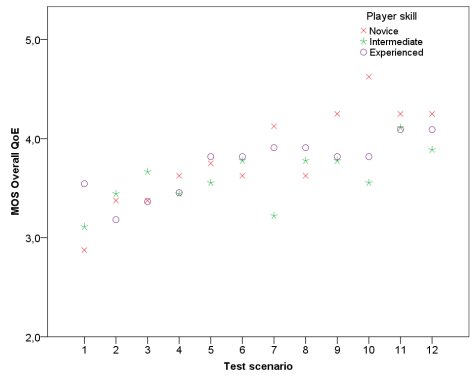
(b) Willingness to continue playing for Study 2

Fig. 8. Willingness to continue playing under different test conditions for tested games

Additionally, we decided to examine the impact of user parameters on QoE, primarily in terms of player's previous gaming experience. Overall QoE scores on a per test scenario basis for SS3 in both studies are shown in Figure 9. Considering only the results from Study 1, experienced players tended to give lower scores for lower quality scenarios, and higher scores for higher quality scenarios, as opposed to novice players. This may be attributed to the hypothesis that novice players are generally less sensitive to different quality variations. As for Study 2, we could not draw a clear conclusion with regards to the impact of player skill on QoE scores as in Study 1. Future tests are needed to further investigate this impact.



(a) QoE scores per skill level for SS3 in Study 1



(b) QoE scores per skill level for SS3 in Study 2

Fig. 9. Mean ratings of overall QoE per skill level for different test scenario (scenarios arranged according to ascending mean QoE)

3.3 Implications of subjective results on game categorization

Based on the subjective results and statistical analysis reported in the previous subsections, we highlight the following main findings of each study:

- Different codec configuration strategies may be applied to different types of games in light of bandwidth availability constraints so as to maximizes player QoE. This was clearly shown using two games belonging to different genres (Study 1).
- In certain cases, the **same** codec configuration strategy may be applied to games belonging to different genres (shown in Study 2).

The summarized conflicting findings lead to the conclusion that for deriving different codec (re)configuration strategies, current game classification approaches are not suitable and that there is a need for a novel categorization of games beyond typically used differentiation such as game genre. We thus study the possibility of categorizing cloud games based on objective video metrics (computed from game streams) and degree of player interactivity (in terms of actions per minute). In the next section we describe a broad analysis of objective metrics collected across 25 different games, leading to the proposal of **two distinct clusters of games**. Drawing back on our previously conducted subjective studies, we will see that SS3 and OMD fall within the same cluster (hence the same codec configuration strategy may be applied), while HS falls in a different cluster.

4 VIDEO GAME CATEGORIZATION FOR CLOUD GAMING

4.1 Video characterisation

To empirically quantify the differences between the tested games, we analysed both temporal and spatial characteristics of their video streams. The first set of metrics is extracted according to ITU-T recommendation P.910 (4/2008): Spatial perceptual information (SI) and Temporal perceptual information (TI) [ITU-T Recommendation 1999].

SI is derived based on the Sobel filter. Each video frame (luminance plane) at time n (F_n) is first filtered with the Sobel filter [$Sobel(F_n)$]. The standard deviation over the pixels (std_{space}) in each Sobel-filtered frame is then computed. This operation is repeated for each frame in the video sequence and results in a time series of spatial information of the scene. The maximum value in the time series (max_{time}) is chosen to represent the spatial information content of the scene. This process can be represented in equation form as:

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\} \tag{2}$$

More details in the frame will result in higher values of SI.

TI is based upon the motion difference feature, $M_n(i, j)$, which is the difference between the pixel values (of the luminance plane) at the same location in space but at successive times or frames. $M_n(i, j)$ as a function of time (n) is defined as:

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j) \tag{3}$$

$F_n(i, j)$ is the luminance value of the pixel at the ith row and jth column of the nth frame in time. The measure of temporal information (TI) is computed as the maximum over time (max_{time}) of the standard deviation over space (std_{space}) of $M_n(i, j)$ over all i and j.

$$TI = max_{time}\{std_{space}[M_n(i, j)]\} \tag{4}$$

More motion in adjacent frames will result in higher values of TI. For scenes that contain scene cuts, two values may be given: one where the scene cut is included in the temporal information measure, and one where it is excluded from the measurement (in our case no scene cuts were present and

normal game play was recorded). TI and SI metrics have been extracted through predefined Matlab scripts (authored by Savvas Argyropoulos).

The second set of metrics are metrics proposed by Mark Claypool [Claypool 2009]. As reported in the paper, typical video encoding characteristics (e.g., the size of intra-coded macroblocks) could be used to objectively measure video motion and scene complexity. Therefore, the following metrics based on video characteristics have been proposed for measuring video motion and scene complexity: Percentage of Forward/backward or Intra-coded Macroblocks (PFIM) for the temporal aspect of the video (motion in subsequent images), and Intra-coded Block Size (IBS) for the spatial aspect of video (scene complexity).

The logic behind PFIM as a measure of video motion is the following: motion in the videos is correlated with the percentage of encoded macroblocks, i.e., a video with visual changes from frame to frame will have these changes encoded (either by neighbouring blocks or independently of other blocks), while video without visual changes can skip much of the encoding. Therefore, by analyzing video sequence and calculating the aforementioned percentage, it is possible to obtain an accurate estimation of the video motion as perceived by user.

On the other hand, IBS represents a new measure of scene complexity: if the scene is simple, there is not much information to be encoded. As a result, the intra-coded block size will be small. If the scene is complicated, the IBS will be large to contain all the information.

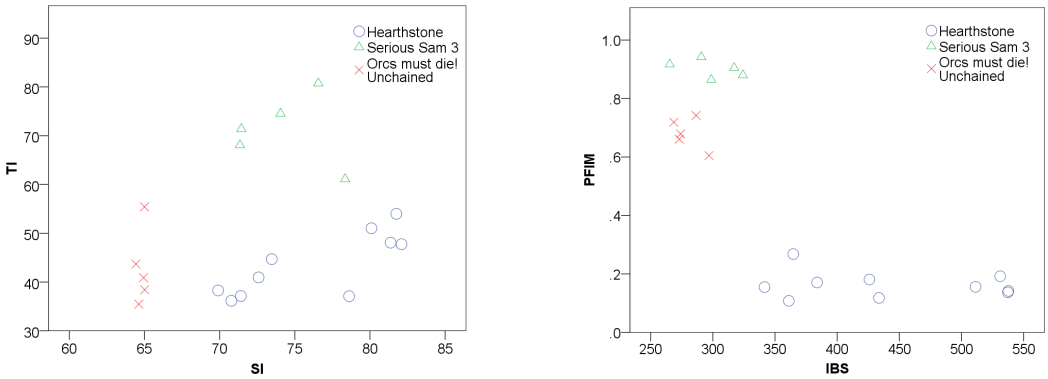
PFIM and IBS metrics were extracted using python scripts created by Mark Claypool [Claypool 2009].

4.2 Preliminary video game traces analysis

In our previous work [Slivar et al. 2016], we reported on the results of video characterization based on objective video metrics using a small cloud gaming video dataset, which implied the possibility of automatic game categorization. This categorization may subsequently be utilized for selecting optimal codec configuration strategies for cloud gaming. Motivated by the results of the conducted user studies that demonstrated that commonly used game classifications (e.g., in terms of graphics detail, game play pace, input rate, etc.) are not necessarily applicable when creating (or identifying the need for) different cloud gaming QoE models, we decided to explore the potential for alternative ways of categorizing games that could be subsequently utilized for selecting optimal video codec configuration strategies.

As an initial step, we opted to check whether there is a difference and/or similarities in video streams between the games tested in conducted user studies, i.e., empirically verify if objective video metrics could indeed be used to differentiate between games. Computed video metrics were plotted and are shown in Figure 10. Each of the dots represents one gaming session played by the same player (at 30 fps and a bitrate of 10 Mbps). The FRAPS application[3] was used to record game play sessions that lasted 30 seconds each. The gaming sessions included distinct game scenes most commonly associated with the tested game's game play. With regards to TI and SI metrics, we can observe that there are indications of clustering behaviour of the video traces for each separate game, but there is a lack of a cluster presence that we expected to be for the video traces of SS3 and OMD based on reported QoE in Study 2. Nevertheless, TI and SI video metrics for SS3 and OMD are more condensed in comparison with HS, indicating that SS3 and OMD game play in terms of game actions is mainly consistent, while for HS it is more dynamic (e.g., choosing cards, waiting for an opponent, playings cards with complex animations, etc.). With respect to PFIM and IBS metrics, both internal and external clustering behaviour can be observed for tested games, as we would expect from this set of video metrics derived for specific purposes of analysing motion and scene

[3]FRAPS, http://fraps.com/



(a) TI and SI scores for tested games

(b) PFIM and IBS scores for tested games

Fig. 10. Scores for different video metric for tested games

complexity of game play videos. Even though there is more spread for the spatial component for HS in comparison with the other two games, the similarity in values of these video metrics can be clearly observed for SS3 and OMD, along with the distinction of these two games and HS (higher temporal and lower spatial scores). These results support our claim that categorization of video games could be achieved by analysing objective video metrics of cloud gaming video stream.

4.3 Methodology

The first step of video game categorization based on objective video metrics for cloud gaming was obtaining a large set of video game traces for analysis purposes. Therefore, we collected video game traces in a laboratory environment slightly different from the one during previous user studies. Similar to the user studies, Valve's Steam In-Home streaming platform was used as the cloud gaming environment. The Steam In-Home Streaming client in this case was installed on a HP Probook 4530s laptop (Windows 10 OS with Intel 2.5 Ghz i5 processor, 4GB RAM and AMD Radeon HD 7400M graphic card), while Steam In-Home Streaming server was installed on a Windows PC desktop (Windows 8 desktop with Intel 3.6 Ghz i7 processor, 8GB RAM and NVIDIA GeForce GTX 970 graphic card). The PC server and laptop client are connected via a wireless access point (both the PC server and the laptop client have a wired connection). The FRAPS application was used once again to record game play sessions. All video traces were recorded at a video encoding frame rate of 30 fps and 10 Mbps video bitrate. Tested games were played in HD-ready resolution (720p) with default graphics settings. For each of the tested games, we recorded between 5-10 game play video traces that lasted exactly 30 seconds each in order to obtain a large enough sample of game play for each game. Alongside game play recording, we also measured the intensity of user interaction, thus collecting mouse and keyboard input during game play by using the Mousotron application [4].

With respect to tested games, we recorded gaming sessions of 25 different video games. During the selection of games, we made sure that the traditional game genres were represented with at least two games. The set of video games for which we recorded game play is shown in Table 3. As a result, we gathered **225 different video traces** that were included in further analysis. For each of the recorded video traces, the following temporal and spatial characteristics were calculated: Spatial perceptual information (SI), Temporal perceptual information (TI),

[4]Mousotron, http://www.blacksunsoftware.com/mousotron.html

Table 3. Selected games for recording video game traces

Game genre	Selected games
Role-playing game	Bastion, Fable II, The Elder Scrolls V: Skyrim, South Park: The Stick of Truth, Orcs Must Die! Unchained
Action game	Batman: Arkham Origins, Joe Danger 2: The Movie, Rocket League
Racing game	Burnout: Paradise, GRID 2, Trials: Evolution
Strategy game	Civilization V, Company of Heroes 2, Medieval II: Total War, Warhammer 40,000:Dawn of War – Dark Crusade
First-person shooter	Counter Strike: Global Offensive, Far Cry 2, Serious Sam 3
Multiplayer online battle arena	DotA 2, Heroes of the Storm
Card game	Hearthstone, Poker Night 2
Other genres	Runner 2, The King of Fighters XIII, Halo: Spartan Assault

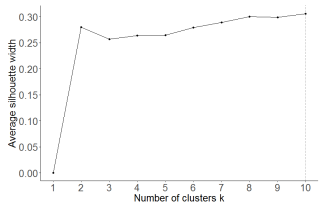
Percentage of Forward/backward or Intra-coded Macroblocks (PFIM) and Intra-coded Block Size (IBS). **The complete data set with annotated video characteristics is available online** at http://www.fer.unizg.hr/qmanic/data_sets.

To investigate the extent to which different playing styles and past player experience may impact the objective video metrics, we performed an additional study with 12 new participants differing in self-reported experience level (3 novice, 3 experienced, and 6 intermediate skilled players). Players recorded game play videos for two representative games: *Heroes of the Storm (HotS)* and *Serious Sam 3 (SS3)*. The results of computing objective video metrics showed that playing style and previous player experience did not have an impact on categorizing games into the found clusters, as will be further explained in the following section.

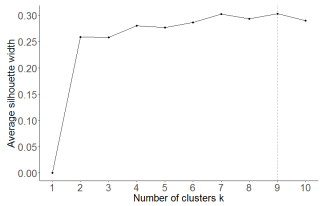
4.4 Results

To identify game categories, we decided to perform cluster analysis on gathered information about recorded video traces. Due to the nature of the problem and type of collected data, we opted for k-means clustering as a clustering technique which is commonly used for performing this type of unsupervised learning tasks. Since k-means clustering requires the number of cluster k as an input parameter, determining the most appropriate number of cluster in the data set was a primary objective. In doing so, we utilized one of the most widely used internal clustering validation measures for k-means clustering, silhouette analysis [Rousseeuw 1987]. Silhouette analysis measures how well an object is clustered (similar with other objects in the same cluster) as compared to other clusters. The silhouette values range from -1 to 1, where a value closer to 1 indicates that the object is well clustered, while a value closer to -1 estimates that the objects should be moved to another cluster. Additionally, due to the low-dimensionality of the data set, we were able to perform a thorough dimension reduction analysis aimed at avoiding excessive sparseness and dissimilarity of the collected video traces.

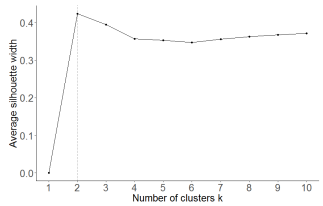
We have limited the silhouette analysis up to a maximum of ten clusters in the data set, along with reducing dimensions of the data set in such a way that the associated metrics were still grouped together. The results of silhouette analysis combined with dimension reduction are presented in Figure 11. Overall, it can be observed that with reducing the number of dimensions, k-means clustering achieves better clustering results, noticeable by the higher values of average silhouette width. When all data dimensions (metrics) are considered, the recommended number of clusters



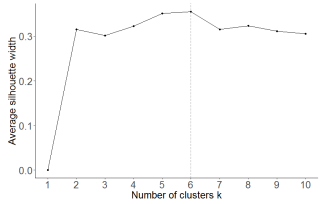
(a) All input metrics (dimensions) considered.



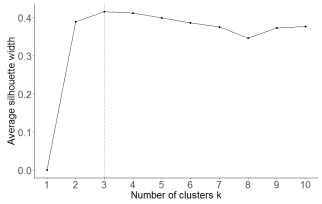
(b) Only TI, SI, and APM metrics considered.



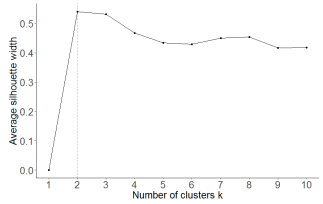
(c) Only TI and SI metrics considered.



(d) Only video metrics TI, SI, PFIM, and IBS considered.



(e) Only PFIM, IBS, and APM metrics considered.



(f) Only PFIM and IBS metrics considered.

Fig. 11. The impact of the chosen number of clusters on the silhouette coefficient for different sets of input metrics. The complete set of collected metrics includes: SI, TI, PFIM, IBS, average actions per minute (APM).

by silhouette analysis is the maximum value of ten clusters, confirming the assumption that a high number dimensions can induce a high dissimilarity in observations, resulting with poor clustering results. As can be seen from the results of the cluster analysis, the APM metric was omitted in the process of reducing the number of dimensions, thus showing that this measure of player interactivity is likely not appropriate in the context of game categorization for cloud gaming. Furthermore, when comparing clustering results with different sets of video metrics, it can be noticed that PFIM and IBS metrics achieve better clustering results than TI and SI metrics, as expected due to the origin of the PFIM and IBS video metrics previously described.

The best clustering results are achieved in the case when only PFIM and IBS video metrics are included in clustering analysis, with average silhouette width higher than 0.5. It should be noted that average silhouette width above 0.5 means that a reasonable clustering structure has been found, while lower values in most cases imply that clustering results could be artificial, as reported in [Kaufman and Rousseeuw 1990]. Therefore, considering clustering options, we have decided to use only PFIM and IBS metrics, and **selected the number of clusters as equal to 2**, which corresponds to the highest average silhouette width across all clustering results. The silhouette plot of the selected data for k-means clustering when k equals 2 is shown in Figure 12. The average silhouette width of the cluster over all data is 0.54, estimating that data is tightly grouped in the clusters. We also note that from a practical cloud gaming service provider point of view, it would likely not make sense to implement a large number of different codec configuration strategies due to the complexity of deriving such strategies using subjective tests, implementation complexity introduced with multiple strategies, as well as cost to benefit ratio.

The results of applying the k-means clustering method with 2 clusters on the collected data set are shown in Table 4. The obtained clusters are highly independent and 80% of the games have 100% of their videos placed in the appropriate cluster. The majority of games with sub 100% accuracy of video cluster placement have values of 80% and 90% of correct cluster placement with 60% being the lowest value. For games with game play videos that are not consistently clustered in the same

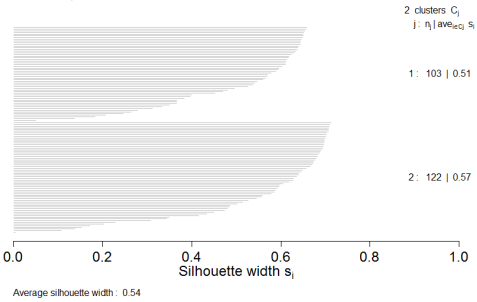


Fig. 12. Silhouette plot of the collected data for k = 2

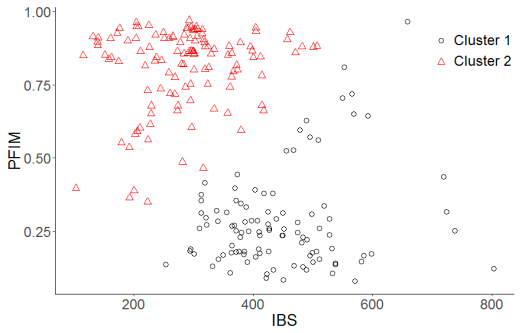


Fig. 13. PFIM and IBS score for the clusters

Table 4. k-means clustering results of video games

Games in Cluster 1	% of video traces grouped in Cluster 1	Games in Cluster 2	% of video traces grouped in Cluster 2
Bastion	90%	Burnout: Paradise	100%
Civilization V	100%	Batman: Arkham Origins	100%
Company of Heroes 2	100%	Counter Strike: GO	100%
DotA 2	90%	Far Cry 2	100%
Fable II	80%	GRID 2	100%
Hearthstone	100%	Halo: Spartan Assault	100%
Heroes of the Storm	100%	Joe Danger 2: The Movie	100%
Medieval II: Total War	100%	Orcs Must Die! Unchained	100%
Poker Night 2	100%	Rocket League	100%
South Park: The Stick of Truth	60%	Runner 2	100%
The King of Fighters XIII	100%	Serious Sam 3	100%
Warhammer 40,000: DoW	100%	The Elder Scrolls V: Skyrim	90%
		Trials: Evolution	100%

cluster (e.g., South Park: The Stick of Truth), an analysis of a larger set of recorded gaming sessions (e.g., 10 videos) is necessary to determine which of the clusters is more fitting for the game. To visualize the obtained clusters, we include a scatter plot of analysed objective video metrics (Figure 13). The cluster centroids are (0.283, 455.711) for the Cluster 1 and (0.786, 284.069) for the Cluster 2. It can be observed that Cluster 2 contains games with high video motion that contains a smaller amount of video information (high PFIM, low IBS). On the other hand, Cluster 1 contains games with low video motion, however when the objects in the screen move, the coding block size is quite large (low PFIM, high IBS).

As previously mentioned, in addition to the 225 video traces used for cluster analysis, additional game play traces were collected by recording players with different self-reported experience levels so as to test the impact of player style and experience level on obtained objective video metrics (and consequently game categorization). All additional players recorded 3 game play traces for two chosen games: HotS (which we chose as a representative game from Cluster 1) and SS3 (chosen as a representative game from Cluster 2). This resulted in an additional 72 video traces. After computing the objective video metrics (PFIM and IBS) for these games, results showed that in all cases HotS

Table 5. Example of video codec configuration strategy in terms of targeted frame rate (video encoding bitrate assumed to be set based on estimated available bandwidth)

Available bandwidth	QoE-driven codec configuration strategy	
	Hearthstone	Serious Sam 3 / Orcs Must Die! U
3 Mbit/s	25 fps	45 fps
5 Mbit/s	25 fps	45 fps
10 Mbit/s	25 fps	60 fps

was categorized into Cluster 1, and SS3 into Cluster 2, thus indicating that previous player's gaming experience and playing style did not have an observable impact on the categorization.

With respect to game genres, we can notice from Table 4 that games that belong to the same traditional video game genre may be clustered into different clusters (e.g., Fable II, Skyrim). Assuming that in this case we employ a video encoding adaptation strategy which applies to all of the games from the same game genre, resulting QoE-driven service adaptation would likely be inefficient for some games. Likewise, we can observe games that are clustered into the same cluster, even though they are not from the same game genre (e.g., SS3 and OMD) and are completely different in terms of game play and camera perspective which are commonly used to differentiate video games. Although the used clustering validation measures (referring to silhouette analysis) suggest that PFIM and IBS give the strongest results in terms of clustering division, we acknowledge that additional user studies need to be performed to further validate the clustering results when comparing more encoding strategies, and to justify the use of PFIM and IBS metrics over other metrics (e.g., SI and TI). Nevertheless, the clustering results clearly demonstrate the possibility of video game categorization based on objective video metrics that could be used for deriving video encoding configuration strategies to achieve high QoE for each categorized video game.

```
Data: game video traces, cluster centroids, video encoding configuration strategies
Result: appropriate video encoding configuration strategy for a game
cluster1, cluster2;
while game video traces do
    calculate PFIM and IBS scores for current;
    calculate distance of PFIM and IBS scores to cluster centroids;
    if distance closer to Cluster 1 centroid then
        cluster1++;
    else
        cluster2++;
    end
end
if cluster1>cluster2 then
    assign Cluster 1 video encoding adaptation strategy;
else
    assign Cluster 2 video encoding adaptation strategy;
end
```

Algorithm 1: Algorithm for choosing appropriate video encoding adaptation strategy

An example derived based on our subjective studies is illustrated in Table 5. The illustrated video encoding strategies shown in Table 5 are derived based on our subjective studies and are an example of strategies that utilize a video encoding configuration that achieves the highest QoE score for the considered bandwidth. The video encoding strategy for SS3/OMD employs a video codec configuration that achieves the highest QoE score in Study 2 (as Study 2 has considered both SS3 and OMD). For SS3, both studies indicate that different video encoding configurations

should be employed for a different bitrate. So in that case one option could be to additionally utilize willingness to play results to determine a final codec configuration for SS3. However, clearly the best option would be to conduct further user studies to obtain more reliable results. Also, given that the cloud gaming server does not have unlimited system resources at its disposal, employing a codec configuration that uses the least amount of system resources could be considered more appropriate. Therefore, for Hearthstone, changing of the frame rate is not required (constant 25 fps), as the perceived QoE remained the same for all test scenarios.

We clarify the process of assigning appropriate video encoding adaptation strategies to games belonging to different clusters in Algorithm 1. The algorithm illustrates the process of categorizing a new game, given multiple video traces of the game play and previously found clusters. The actual adaptation strategies are derived based on the results of subjective studies.

5 CONCLUSIONS

While cloud gaming represents a promising paradigm shift in the domain of online gaming, challenges arise in meeting the strict bandwidth and delay requirements of game streaming. Such challenges inherently call for codec configuration and adaptation strategies capable of meeting strict player QoE requirements and adapting the service to variable network conditions. Given the wide diversity of games and their corresponding QoE requirements, it is clear that different strategies may be applied to different categories of games. The open question that we address is how can such categories potentially be derived.

The contributions of this paper can be summarized as threefold. First, two subjective laboratory studies have been conducted which provide valuable insights into the impacts of different codec configuration strategies on three different games. Secondly, using as a basis the subjective results, we propose a novel game categorization based on objective video metrics that could be utilized for deriving appropriate video encoding configuration strategies for cloud gaming. The results have shown that the game type tested clearly needs to be taken into account when performing server-side QoE-driven service adaptation. Moreover, existing game classifications (e.g., game genres) are not necessarily applicable for differentiating video games in the context of choosing appropriate video encoding configuration strategies. Finally, a third contribution of the paper is a large and openly available dataset of 225 different game play videos recorded across 25 different games and annotated with objective video metrics. Based on the analysis of collected game play video traces, objective video metrics of game play videos combined with player actions intensity were used as the input for clustering games into two clusters. The proposed game categorization could then be used for determining adaptation strategies for clusters of games, which could in the future automate the process of deciding on the best video encoding strategy for a particular game, alleviating the need to conduct subjective studies for additionally considered (or newly emerging) games.

Future studies will aim to validate the proposed game categorization for cloud gaming using a larger game data set. Moreover, we aim to verify the clustering results by conducting additional subjective studies testing games both from different clusters and from the same cluster. Finally, we would like to test the impact of different encoding adaptation strategies on client device battery consumption, as additional potential motivation to adapt codec configuration parameters.

REFERENCES

Hamed Ahmadi, Saman Zad Tootaghaj, Mahmoud Reza Hashemi, and Shervin Shirmohammadi. 2014. A game attention model for efficient bit rate allocation in cloud gaming. *Multimedia Systems* 20, 5 (2014), 485–501.

Justus Beyer, Richard Varbelow, Jan-Niklas Antons, and Sebastian Möller. 2015. Using electroencephalography and subjective self-assessment to measure the influence of quality variations in cloud gaming. In *Quality of Multimedia Experience*

(QoMEX), 2015 Seventh International Workshop on. 1–6.

Wei Cai, Ryan Shea, Chun-Ying Huang, Kuan-Ta Chen, Jiangchuan Liu, Victor CM Leung, and Cheng-Hsin Hsu. 2016. A Survey on Cloud Gaming: Future of Computer Games. *IEEE Access* 4 (2016), 7605–7620.

Mark Claypool. 2009. Motion and scene complexity for streaming video games. In *Proceedings of the 4th International Conference on Foundations of Digital Games*. ACM, 34–41.

Mark Claypool and Kajal Claypool. 2006. Latency and player actions in online games. *Commun. ACM* 49, 11 (Nov. 2006), 40–45. https://doi.org/10.1145/1167838.1167860

Mark Claypool and Kajal Claypool. 2009. Perspectives, frame rates and resolutions: it's all in the game. In *Proceedings of the 4th International Conference on Foundations of Digital Games*. ACM, 42–49.

Mark Claypool, Jonathan Decelle, Gabriel Hall, and Lindsay O'Donnell. 2015. Surrender at 20? Matchmaking in League of Legends. In *Games Entertainment Media Conference (GEM), 2015 IEEE*. IEEE, 1–4.

Mark Claypool and David Finkel. 2014. The effects of latency on player performance in cloud-based games. In *Proc. of 13th Annual IEEE/ACM NetGames*. 1–6.

Victor Clincy and Brandon Wilgor. 2013. Subjective Evaluation of Latency and Packet Loss in a Cloud-Based Game. In *Information Technology: New Generations (ITNG), 2013 Tenth International Conference on*. IEEE, 473–476.

Matthias Dick, Oliver Wellnitz, and Lars Wolf. 2005. Analysis of factors affecting players' performance and perception in multiplayer games. In *Proc. of 4th ACM SIGCOMM workshop on Network and system support for games (NetGames)*. 1–7.

Hua-Jun Hong, Chih-Fan Hsu, Tsung-Han Tsai, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2015. Enabling Adaptive Cloud Gaming in an Open-Source Cloud Gaming Platform. *IEEE Transactions on Circuits and Systems for Video Technology* PP, 99 (2015), 1–14.

Tobias Hoßfeld, Poul E Heegaard, Martín Varela, and Sebastian Möller. 2016. QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS. *Quality and User Experience* 1, 1 (2016), 2.

Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not enough!. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*. IEEE, 131–136.

P.910 ITU-T Recommendation. 1999. Subjective video quality assessment methods for multimedia applications. (1999).

Michael Jarschel, Daniel Schlosser, Sven Scheuring, and Tobias Hoßfeld. 2013. Gaming in the clouds: QoE and the users' perspective. *Mathematical and Computer Modelling* 57, 11 (2013), 2883–2894.

Leonard Kaufman and Peter J. Rousseeuw. 1990. *Finding groups in data : an introduction to cluster analysis*. Wiley, New York. http://opac.inria.fr/record=b1087461 A Wiley-Interscience publication.

Yeng-Ting Lee, Kuan-Ta Chen, Han-I Su, and Chin-Laung Lei. 2012. Are all games equally cloud-gaming-friendly? An electromyographic approach. In *Network and Systems Support for Games (NetGames), 11th Annual Workshop on*. 1–6.

Yao Liu, Shaoxuan Wang, and Sujit Dey. 2014. Content-Aware Modeling and Enhancing User Experience in Cloud Mobile Rendering and Streaming. *IEEE J. Emerg. Sel. Topics Circuits Syst.* 4, 1 (2014), 43–56.

Sebastian Möller, Dennis Pommer, Justus Beyer, and Jannis Rake-Revelant. 2013. Factors Influencing Gaming QoE: Lessons Learned from the Evaluation of Cloud Gaming Services. In *Proceedings of the 4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. 1–5.

Peter Quax, Anastasiia Beznosyk, Wouter Vanmontfort, Ronald Marx, and Wim Lamotte. 2013. An evaluation of the impact of game genre on user experience in cloud gaming. In *Games Innovation Conference (IGIC), 2013 IEEE Intl.* 216–221.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.

Ivan Slivar, Lea Skorin-Kapov, and Mirko Suznjevic. 2016. Cloud gaming QoE models for deriving video encoding adaptation strategies. In *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 18.

Ivan Slivar, Mirko Suznjevic, and Lea Skorin-Kapov. 2014. Empirical QoE Study of In-Home Streaming of Online Games. In *Proceedings of the 14th Annual Workshop on Network and Systems Support for Games, NetGames*. Nagoya, Japan.

Ivan Slivar, Mirko Suznjevic, and Lea Skorin-Kapov. 2015. The impact of video encoding parameters and game type on QoE for cloud gaming: A case study using the Steam platform. In *Quality of Multimedia Experience (QoMEX), 2015 7th Intl. Workshop on*. 1–6.

Mirko Suznjevic, Ognjen Dobrijevic, and Maja Matijasevic. 2009. MMORPG Player actions: Network performance, session patterns and latency requirements analysis. *Multimedia Tools and Applications* 45, 1-3 (2009), 191–241.

Mirko Suznjevic, Lea Skorin-Kapov, and Maja Matijasevic. 2013. Impact of User, System, and Context factors on Gaming QoE: a Case Study Involving MMORPGs. In *Proc. of the 12th ACM SIGCOMM Workshop on Network and System Support for Games*.

Shaoxuan Wang and Sujit Dey. 2009. Modeling and characterizing user experience in a cloud server based mobile gaming approach. In *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*. IEEE, 1–7.

Zi-Yi Wen and Hsu-Feng Hsiao. 2014. QoE-driven performance analysis of cloud gaming services. In *Multimedia Signal Processing (MMSP), 2014 IEEE 16th International Workshop on*. IEEE, 1–6.