Received 16 July 2018; | Revised 17 Sept 2018; | Accepted 29 Oct 2018

DOI: xxx/xxxx

RESEARCH ARTICLE

An Objective and Subjective Quality Assessment Study of Passive Gaming Video Streaming

Nabajeet Barman[1] | Saman Zadtootaghaj[2] | Steven Schmidt[3] | Maria G. Martini[1] | Sebastian Möller[3]

[1]Kingston University, London, United Kingdom
[2]Deutsche Telekom, Berlin, Germany
[3]Technische Universität Berlin, Berlin, Germany

Correspondence
Nabajeet Barman,
Department of Computer Science and Mathematics,
Kingston University,
London, United Kingdom KT1 2EE.
Email: nabajeetbarman4@gmail.com

Abstract

Passive gaming video streaming applications have recently gained much attention as evident with the rising popularity of many Over The Top (OTT) providers such as Twitch.tv and YouTubeGaming. For the continued success of such services, it is imperative that the user Quality of Experience (QoE) remains high which is usually assessed using subjective and objective video quality assessment methods. Recent years have seen tremendous advancement in the field of objective video quality assessment (VQA) metrics, with the development of models that can predict the quality of the videos streamed over the Internet. A study on the performance of objective VQA on gaming videos, which are artificial and synthetic and have different streaming requirements than traditionally streamed videos, is still missing. Towards this end, we present in this paper an objective and subjective quality assessment study on gaming videos considering passive streaming applications. Subjective ratings are obtained for 90 stimuli generated by encoding six different video games in multiple resolution-bitrate pairs. Objective quality performance evaluation considering eight widely used VQA metrics is performed using the subjective test results and on a dataset of 24 reference videos and 576 compressed sequences obtained by encoding them in 24 resolution-bitrate pairs. Our results indicate that VMAF predicts subjective video quality ratings the best, while NIQE turns out to be a promising alternative as a no-reference metric in some scenarios.

KEYWORDS:
Multimedia Services, Gaming Video Streaming, Quality Assessment, Quality of Experience

1 | INTRODUCTION

Gaming video streaming applications are becoming increasingly popular. They can be divided into two different, but related, applications: interactive and passive. Interactive gaming video streaming applications are commonly known as cloud gaming where the game is rendered on a cloud server. The user receives the streamed gameplay video on a client device, and the user game inputs such as keystrokes are sent back to the server. Such applications have received lots of attention, resulting in the rapid development and acceptance of such services [1]. On the other hand, passive gaming video streaming refers to applications such as Twitch.tv, where viewers can watch the gameplay of other gamers. Such applications have received much less attention from both the gaming and video community despite the fact that Twitch.tv, with approximately two million streamers and over 15 million daily active users, is alone responsible for the 4th highest total traffic during peak hours [2]. With the increasing popularity

FIGURE 1 Screenshots of the gaming videos used in this work

of such services, along with the demand for other over-the-top (OTT) services such as Netflix and YouTube, the demand on network resources has also increased. Therefore, to provide the end-user with a service at a reasonable Quality of Experience (QoE) and satisfy the user expectation of any time, anyplace and any-content video service availability, it is necessary to optimize the video delivery process.

For the assessment of video quality, and hence QoE, typically subjective tests are carried out. However, these tests are time-consuming and expensive. Thus, numerous efforts are being made to predict the video quality through video quality assessment (VQA) metrics. Depending on the availability and the amount of source information, objective video quality assessment (VQA) metrics can be categorized into full-reference (FR), reduced-reference (RR), and no-reference (NR). So far, these metrics have been developed and tested for non-gaming videos, usually considering video on demand (VoD) streaming applications. Also, some of the metrics such as NIQE and BRISQUE are based on qualities inherent to natural images (for details see Section 4.1). Gaming videos, on the other hand, are artificial and synthetic in nature and have different streaming requirements (1-pass, Constant Bitrate (CBR)). Hence, the performance of these VQA metrics remains an open question. Our earlier study[3] found some differences in the performance of such metrics when comparing gaming videos to non-gaming videos.

Towards this end, this paper presents a subjective and objective evaluation of gaming videos considering a passive streaming scenario which is an extension of our previous works[4,5]. In Section 2 we present a dataset and the evaluation methodology. The subjective quality assessment study and the results are presented in Section 3. Section 4 introduces and presents a performance evaluation of the eight most popular and widely used quality metrics in terms of correlation scores. Various results and observations are discussed. Additionally, the results of the metric performance for various pooling strategies and complexity based game classification strategy are presented and discussed. Section 5 finally concludes the paper with a discussion of how the results and observations reported in this study can be used in future studies.

2 | DATASET AND EVALUATION METHODOLOGY

2.1 | Description of the Games

Gaming videos streamed over the Internet cover a wide range of games from different genres of varying degree of encoding complexity. For this study, we recorded 24 video sequences from a total of 12 games (each game two sequences) taking into

TABLE 1 Summary of selected games, respective genre and Twitch.tv ranking (RPG: Role Playing Game; MOBA: Multiplayer Online Battle Arena; MMORPG: Massively Multiplayer Online Role-playing Game)

Game	Genre	Twitch.tv Ranking
Counter Strike Global Offensive (CSGO)	FPS	3
Diablo III (Diablo)	RPG	31
Defense of the Ancients 2 (Dota2)	MOBA	4
FIFA 2017 (FIFA)	Sports	18
H1Z1: Just Survive (H1Z1)	Survival	53
Hearthstone (HS)	Collectible Card Game	8
Heroes of the Storm (HoTS)	MOBA	21
League of Legends (LoL)	MOBA	1
Project Cars (PC)	Racing Simulator	100+
PlayerUnknown's Battleground (PUBG)	Battle Royale	5
Startcraft 2 (SC)	Strategy	25
World of Warcraft (WoW)	MMORPG	9

account the genre, popularity (number of viewers on Twitch.tv) and video encoding complexity, as shown in Figure 1 and summarized in the Table 1 .

2.2 | Source Videos

The video sequences were captured losslessly in the RGB format at 30 frames per second (fps) using FRAPS[6] (version 3.5.99). For subjective video quality tests using non-gaming video content, typically a very short stimulus duration of only 10-15 seconds is used. However, as presented and discussed by the authors in various works[7,8,9], such a short duration may not be sufficient to cover a representative scene of a game, as the complexity over the length of the video sequence may vary drastically due to player behavior. Therefore, for this study, in line with the earlier recommendations, we selected a stimulus duration of 30 seconds. The scenarios for the captured sequences were chosen in such a way that they represent a common player behavior and are also representative of the actual game characteristics as usually streamed by the OTT services and watched by viewers. The captured sequences were then processed and converted into YUV format using FFmpeg.

As discussed earlier, for selection of game as well as for the selection of the respective video sequences, the video complexity of recorded sequences was taken into account. Spatial information (SI) and temporal information (TI) values as defined in ITU-T Rec. P.910 can be used as an approximate measure of complexity, with high SI and high TI values representing a high level of complexity. From Figure 2 , it can be observed that SI and TI values do not vary much for video sequences for the same game considering the fact that they represent different scenarios from the game (for some cases, even different levels).

In order to further analyze the temporal and spatial behavior of recorded video sequences of our dataset, we calculate and plot the variation of SI and TI over the total duration of the video sequences (900 frames) instead of using single maximum value, as defined originally by ITU-T Rec. P.910[10]. Figure 3 presents the Box Plot considering the individual frame level SI and TI values of the games. It can be observed from Figure 3 that there is a small variation of SI and TI over the duration of the video for games with omnipresent perspectives (players view and simultaneously influence the entire set of resources under their control[7]), such as Dota2, LoL and SC. In addition, for games with the first-person perspective (the camera location is synonymous with the avatar's eyes and the game world objects appear smaller and closer together the farther they are from the camera location[7]) such as CSGO and H1Z1 high variation of SI and TI is observed. A comprehensive analysis of video game complexity has been carried out in[11] which was used as a reference in the selection of video sequences for the subjective quality assessment as discussed later in Section 3.1.

2.3 | Encoding Settings

For streaming video games over the Internet, usually Constant Bitrate (CBR) is used as a rate control mode. CBR is usually selected due to the inherent feature of video games that can result in highly dynamic scenes followed by dull moments of

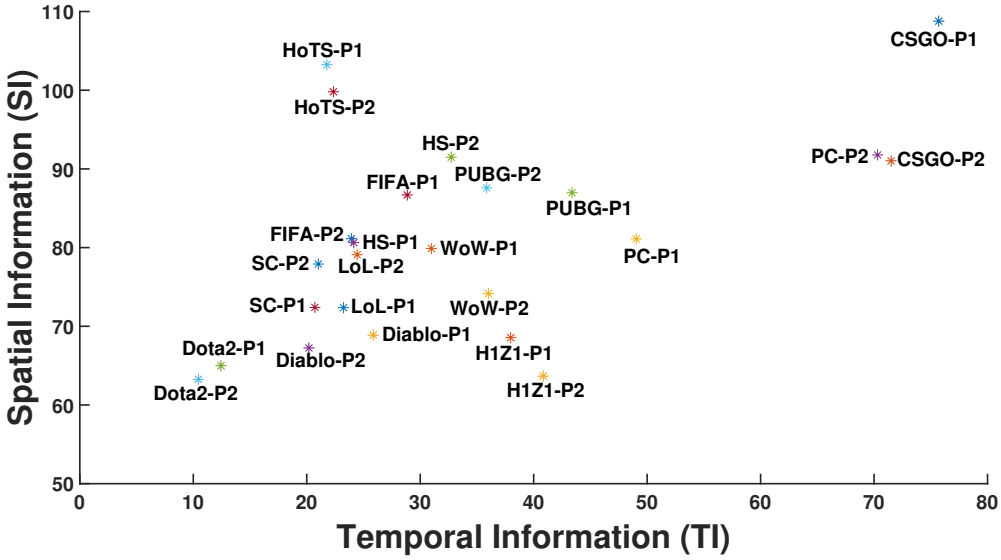

FIGURE 2 SI and TI plot for the 24 gaming video sequences.

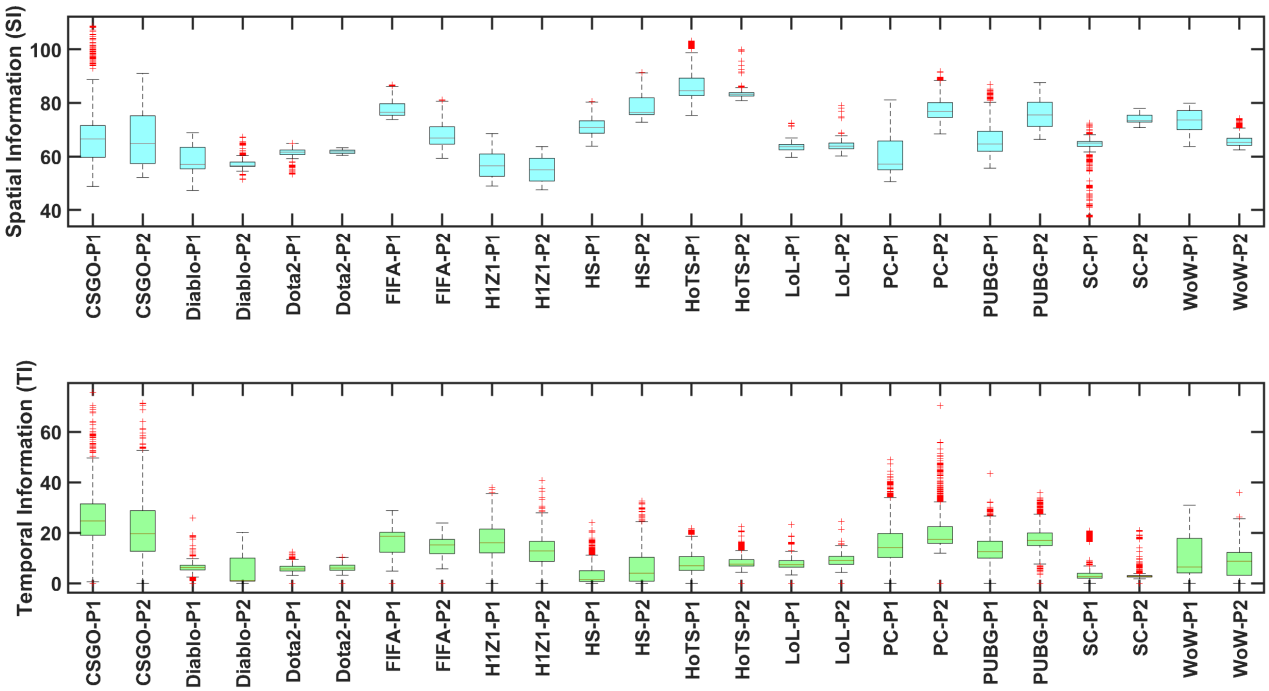

FIGURE 3 Box plot of SI (top) and TI (bottom) values for short duration gaming video sequences.

gameplay. Using other rate control modes of encoding such Variable Bitrate (VBR), can result in stalling of the video playback at the end-user when high dynamic scenes appear after long dull moments of gameplay, leading to lower QoE[12]. Based on the discussion provided by Barman and Martini[13], we select the encoding settings based on the recommendations by various OTT

TABLE 2 Summary of video encoding parameters.

Parameter	Value
Duration	30 sec
Resolution	1080p, 720p, 480p
Frame Rate	30
Encoder	FFmpeg
Encoding Mode	CBR
Video Compression Standard	H.264, Main 4.0
Preset	veryfast

TABLE 3 Resolution-Bitrate pairs used to obtain distorted (compressed) video sequences. The bitrates in bold text refer to the bitrates used in the subjective quality assessment.

Resolution	Bitrate (kbps)
1080p	**600**, **750**, 1000, **1200**, 1500, **2000**, 3000, **4000**
720p	**500**, **600**, 750, 900, **1200**, 1600, **2000**, 2500, **4000**
480p	**300**, 400, **600**, 900, **1200**, **2000**, **4000**

service providers. Our preliminary studies (not reported here) have shown no observable difference in the performance of VQA metrics when considering other encoding settings (VBR, 2-pass). However, in line with the industry wide used settings and recommendations, we choose CBR and 1-pass encoding settings as summarized in Table 2 . Table 3 describes the resolution-bitrate pairs considered in this work.

3 | SUBJECTIVE ASSESSMENT

Subjective tests are means to measure and assess the QoE of a multimedia service. To assess the video quality of the encoded gaming video sequences, we carried out a subjective test in which ratings for different gaming videos encoded at different resolutions and bitrates are obtained from human observers. Based on the objectives of the test, there exists many different possible test settings and methodologies that one can select from, which are summarized in ITU-R Rec. BT.500[14] and ITU-T Rec. P.910[14]. We discuss next the subjective test settings and methodology used in this work.

3.1 | Test Environment and Set up

The subjective test was conducted at the Berlin Institute of Technology in standardized test room according to ITU-R Rec. BT.500[14]. In order to avoid test participant fatigue, we limit the test duration by selecting six video sequences, three resolutions and five bitrates at each resolution. The resolution-bitrate pairs were chosen with the aim to cover a broad range of quality degradations without reaching saturation. The selection of video games was made after an in-depth analysis of video content (genre, popularity etc.) and considering game video complexity classification presented by Zadtootaghaj *et al.*[11]. The selected six games are as follows: CSGO and H1Z1 (high complexity), FIFA, LoL, and PC (medium complexity) and HS (low complexity). Table 3 presents the selected resolution-bitrate pairs used in this study and with the resolution-bitrates pairs used for the subjective assessment highlighted in bold.

In order to avoid any unexpected artifacts of re-scaling of the videos by the video player, the downscaled, encoded MP4 video sequences at 720p and 480p resolutions were decoded and rescaled to 1080p, YUV videos using the *bilinear* scaling filter. The decoded raw YUV videos were then put in an *.avi* container for playback on a 24″ ViewSonic display monitor using the VLC player[15]. The order of resolution-bitrate pairs, as well as the order of the games, were randomized to avoid learning effects. A training session was conducted prior to the test in order to get users familiar with the test set up and the interface of the tool. For the training session, four video sequences from two games (Diablo and WoW, which are different from the ones considered for the subjective tests) were used. The visual acuity and color blindness of all participants were checked by using Snellen charts and Ishihara plates, respectively. The subjective ratings of the test participants who did not fulfill the visual capability

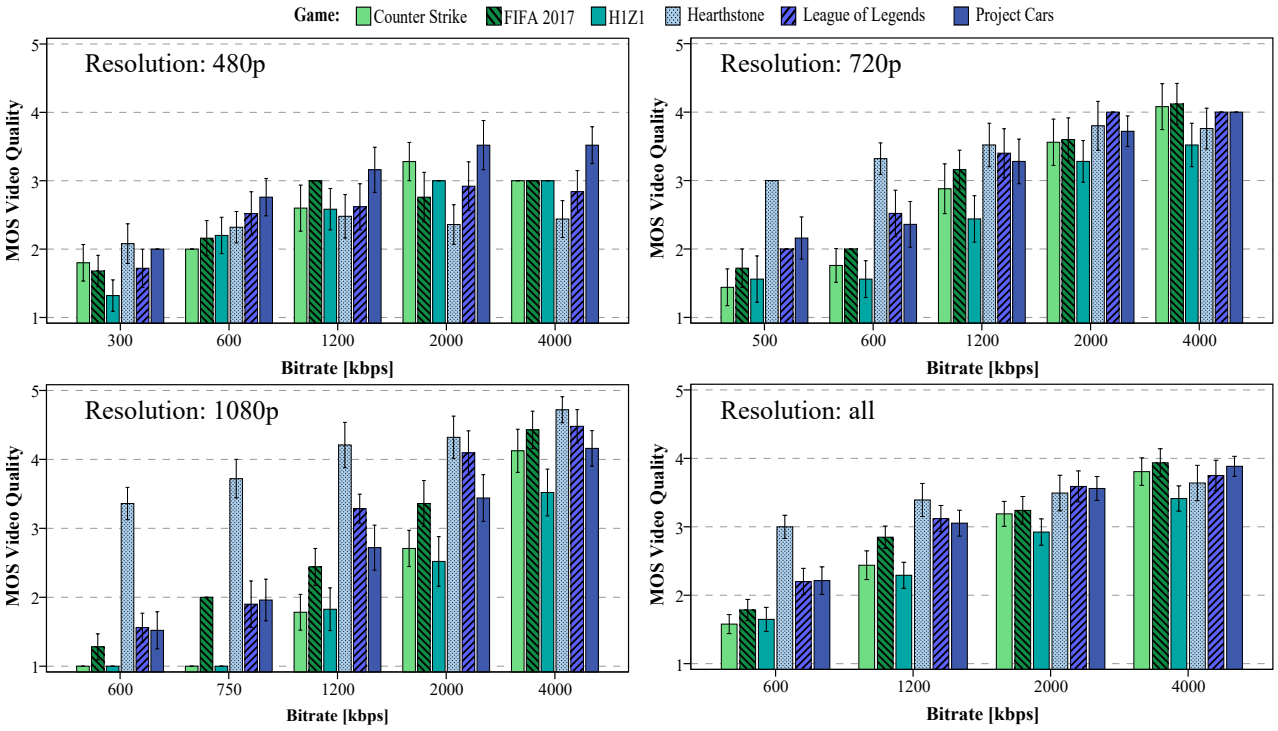

FIGURE 4 Barplots of video quality ratings for the six selected games for the used bitrates at different resolutions.

requirements were removed from the dataset. In the end, the dataset consisted of the subjective scores of a total of 25 subjects with a median age of 29 years. For use cases such as codec evaluation, quality metric performance evaluation and comparison, cross-lab validation studies, etc., subjective video quality ratings are of very high interest to the research community. Hence, we make available the results as well as the reference and distorted videos as an open source dataset, GamingVideoSET[5].

3.2 | Subjective Scores

Figure 4 presents the bar plot of the video quality ratings in terms of Mean Opinion Scores (MOS) for the six gaming video sequences at 480p, 720p, and 1080p resolutions as well as for all resolutions combined. Based on the Figure 4 , the impact of video complexity is apparent since for high complex games, H1Z1 and CSGO, at 1080p resolution and 600 kbps, no participant rated the video quality better than "bad" (1) while it does not hold true for other games. In addition, it can be observed that at 480p resolution, the video quality gets saturated for bitrates equal and higher than 2 Mbps. Furthermore, the video quality ratings for low complexity game HS are significantly higher than that for the other games at low bitrates for resolutions of 720p and 1080p. Lastly, the ratings of H1Z1 even for 4000 kbps at a resolution of 1080p, are lower than that of the other games.

Along with the subjective opinion ratings measured on a 5-point Absolute Category Rating (ACR) scale, the acceptance of the respective stimulus on a binary scale was also measured. A heat map of acceptance rate created based on the percentage of acceptance at each quality level (ranging from 1 to 5, regardless of condition) for six gaming video sequences used in the subjective test is presented in Figure 5 . It can be observed that for low quality ratings, the acceptance rate for HS was higher than that for the other games and that the acceptance rate for PC for the highest score is higher than that for the other games. These differences might be caused by user factors (e.g., game preferences).

4 | OBJECTIVE ASSESSMENT

4.1 | VQA Metrics

Depending on the amount of reference information required by the VQA metric, they can primarily be divided into three categories: Full Reference (FR), Reduced Reference (RR) and No Reference (NR). While FR metrics typically offer more accurate

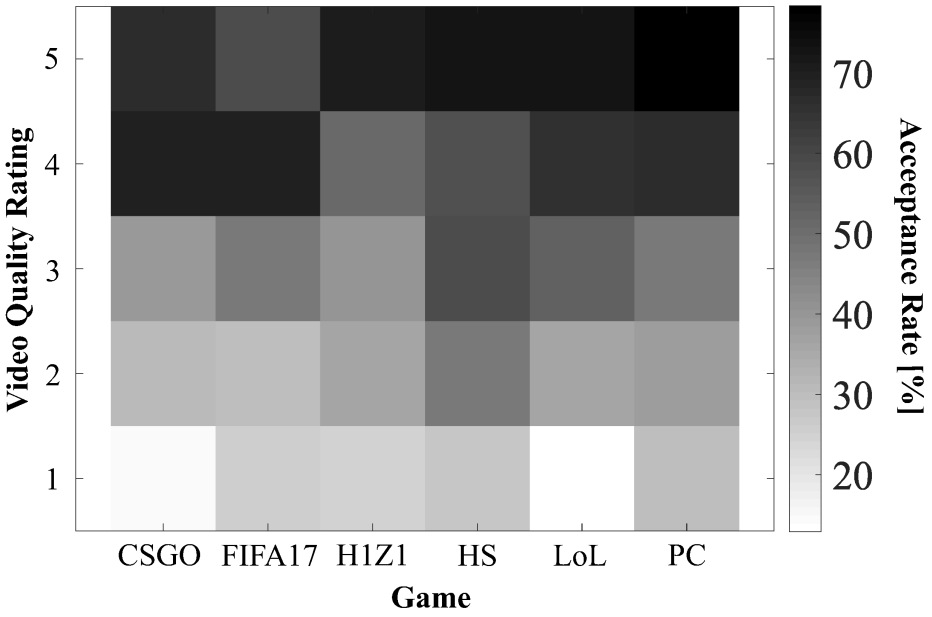

FIGURE 5 Heat map of acceptance rate at each video quality level (regardless of condition) for six gaming video sequences.

video quality predictions, as they have access to more information, they cannot usually be used in many practical use cases (such as live streaming including gaming video streaming). For this reason, the development and evaluation of RR metrics and NR metrics are of high interest. For completion and possible comparative performance evaluation of the VQA metrics from the three different categories, we selected a total of eight most popular and widely used VQA metrics which are discussed next.

4.1.1 | FR metrics

FR metrics refer to VQA metrics which requires the availability of full reference information. For a meaningful comparison within the FR metrics, but also for the comparison with other metric types and subjective ratings, we selected three widely used FR metrics: Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Metric (SSIM)[16] and Video Multi-Method Assessment Fusion (VMAF)[17]. Due to its simplicity and ease of computation, PSNR is one of the most widely used metrics for both image and video quality assessment. SSIM, which computes the structural similarity between the two images, was shown to correlate better with subjective judgment and hence is also widely used for both image and video quality assessment[16]. The PSNR and SSIM values for each frame are temporally pooled (usually averaged) over the video duration to obtain a single score. VMAF is a fusion based metric which combines scores from three different metrics to obtain a single score between 0 to 100, with a higher score denoting a higher quality. The choice of VMAF along with PSNR and SSIM is influenced by our previous work, where we observed that VMAF predictions have a very high correlation with subjective video quality rating[3].

4.1.2 | RR Metrics

Reduced-reference metrics are used when only partial information about the reference video is available. Thus, they are usually less accurate than FR metrics but are useful in applications where there is limited source information available due to scenarios such as limited bandwidth transmissions. As a choice of RR metric, we consider Spatio-temporal-reduced reference entropic differences (ST-RRED)[18] as it is one of the most widely used RR metrics with very good performance on various VQA databases[19]. It measures the amount of spatial and temporal information differences in terms of wavelet coefficients of the frames and frame differences between the distorted and received videos. In this work we use the recently developed optimized version of ST-RRED, known as STRREDOpt, which calculates only the desired sub-band, resulting in almost the same performance as ST-RRED but almost ten times computationally faster[20]. In addition, we also use the recently proposed spatial efficient entropic

differencing for quality assessment (SpEEDQA) model, which is almost 70 times faster than the original implementation of ST-RRED and seven times faster than STRREDOpt as it considers only the spatial domain for its computation[21]. For both these metrics, we used the default settings and implementation as provided by the authors.

4.1.3 | NR Metrics

NR metrics try to predict the quality without using any source information. Since for gaming applications, a high-quality reference video is typically not available, the development of good performing no-reference metrics is of very high importance. For this work, we selected three NR metrics. Amongst them is Blind/referenceless image spatial quality evaluator (BRISQUE)[22], which tries to quantify the possible loss of naturalness in an image by using the locally normalized luminance coefficients. Additionally, we selected the Blind image quality index (BIQI). BIQI is a modular NR metric based on distortion image statistics which is based on natural scene statistics (NSS)[23]. Lastly, we included Natural Image Quality Evaluator (NIQE)[24], a learning-based NR quality estimation metric which uses statistical features based on the space domain NSS model.

4.2 | Objective Metric Performance Evaluation

The aim of a VQA metric is to use objective measurements such as signal fidelity measurements to predict visual quality as perceived by human observers. To evaluate the performance of a VQA metric, typically the correlation between the objective metric score with subjective scores are analyzed. Typically, Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SROCC) describing the relationship between the objective metric score with subjective scores are used. While the Spearman correlation indicates the strength and direction of the monotonic relationship between both scores, the Pearson's correlation determines the strength and direction of the linear relationship between them. A correlation value of 1 indicates a very high positive correlation while a value of minus 1 indicates a strong negative correlation. A value of zero indicates no correlation.

In this work, we measure the performance of the objective metrics in two phases. In the first phase, we compare the performance of the VQA metrics with subjective scores considering the subjective dataset. In the second phase, for a comprehensive evaluation of the VQA metrics on the full dataset, we compare the VQA metric performance with a benchmark VQA metric. Since the encoded videos available are MP4, for FR and RR metric calculations, we instead use the decoded, raw YUV videos obtained from the encoded MP4 videos. The videos at 480p and 720p resolution were rescaled to 1080p YUV format using *bilinear* scaling filter as was done for subjective quality assessment. For NR metric calculations we use the encoded videos at their original resolution (without scaling 480p and 720p videos to 1080p). PSNR and SSIM calculation were done using the VQMT tool available in[25] while for VMAF (version: VMAF_VF0.2.4b-0.6.1) we used the Linux based implementation in[17]. For ST-RREDOpt, SpEED-QA and BIQI we used the implementation made available by the authors using the default settings. NIQE and BRISQUE calculations were done using the inbuilt MATLAB function (version: R2017b)[26].

4.2.1 | Comparison of VQA metrics with MOS

Table 4 shows the correlation values of the eight VQA metrics with respect to MOS scores. The results are reported separately for each resolution as well as all three resolution-bitrate combined (all data). It can be observed that VMAF results in the highest performance in terms of both PLCC and SROCC values across all three resolutions and all data except for SROCC value at 480p. The two RR metrics have a similar performance in terms of correlation values across all resolution-bitrate pairs and over all data. Hence for applications where an increased speed of computation is of high importance, SpEEDQA can be selected as RR metric as it is almost seven times faster than ST-RREDOpt. Among the NR metrics, NIQE performs the best. For 1080p, BIQI and NIQE result in almost similar correlation values. For other resolutions and *all data*, BRISQUE and BIQI perform very similar.

4.2.2 | Impact of resolution on VQA metrics

It can be observed that in general, the performance of the VQA metrics varies across different resolutions. For the FR and NR metrics, the performance of the metrics for higher resolution videos is significantly better than that for lower resolution videos. In contrast, both RR metrics resulted in higher correlation in terms of PLCC with MOS scores for 720p resolution videos, followed by 1080p and 480p resolution videos. Fisher's Z-test to assess the significance of the difference between two correlation coefficients indicates that the difference between 720p and 1080p is not statistically significant, while the difference

TABLE 4 Comparison of the performance of the VQA metric scores with MOS ratings in terms of PLCC and SROCC values. All Data refers to the combined data of all three resolution-bitrate pairs. The best performing metric is shown in bold.

Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.67	0.64	0.80	0.78	0.86	0.87	0.74	0.74
	SSIM	0.58	0.44	0.81	0.78	0.86	0.90	0.80	0.80
	VMAF	**0.81**	0.74	**0.95**	**0.94**	**0.97**	**0.96**	**0.87**	**0.87**
RR Metrics	STRREDOpt	-0.67	-0.56	-0.86	-0.89	-0.83	-0.96	-0.75	-0.77
	SpEEDQA	-0.69	-0.55	-0.88	-0.90	-0.80	-0.96	-0.75	-0.77
NR Metrics	BRISQUE	-0.40	-0.37	-0.76	-0.80	-0.79	-0.76	-0.44	-0.46
	BIQI	-0.41	-0.35	-0.72	-0.70	-0.83	-0.82	-0.42	-0.45
	NIQE	-0.77	**-0.76**	-0.77	-0.73	-0.84	-0.84	-0.72	-0.71

TABLE 5 Comparison of the performance of the VQA metric scores with VMAF scores in terms of PLCC and SROCC values. All Data refers to the combined data of all three resolution-bitrate pairs. The best performing metric is shown in bold.

Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.62	0.60	0.79	0.77	**0.91**	0.92	**0.87**	**0.87**
	SSIM	0.56	0.57	0.69	0.70	0.81	0.84	0.71	0.74
RR Metrics	STRREDOpt	-0.70	-0.88	-0.79	-0.92	-0.80	-0.93	-0.56	-0.63
	SpEEDQA	-0.71	**-0.90**	**-0.80**	**-0.95**	-0.79	**-0.95**	-0.58	-0.65
NR Metrics	BRISQUE	-0.62	-0.62	-0.75	-0.74	-0.74	-0.75	-0.11	-0.11
	BIQI	-0.55	-0.51	-0.70	-0.70	-0.67	-0.68	-0.04	-0.04
	NIQE	**-0.75**	-0.77	-0.79	-0.79	-0.77	-0.75	-0.41	-0.42

between 720p and 480p is significant, $Z = 2.954$, $p < 0.01$. For all eight VQA metrics, the performance for the 480p resolution (cf. Table 4) is considerably lower compared to the same VQA metric performance for the 720p and 1080p resolutions. Also, the decrease in performance for some metrics is higher than others. In figure 6 b this observation is illustrated using PSNR and VMAF as an example. It can be seen that PSNR for different bitrates at 480p resolution is not able to capture the variation in MOS (cf. Figure 6 a) as its values for the 480p resolution almost remain constant even at higher bitrates. VMAF, on the other hand, as evident from Figure 6 c, captures this variation quite well and hence results in increased performance overall and also across each individual resolutions.

4.2.3 | Comparison of VQA metrics with VMAF

In the previous section, we presented and evaluated the performance of the eight VQA metrics based on the subjective ratings using six reference gaming video sequences and 15 resolution-bitrate pairs. It was found that across all conditions, VMAF resulted in the highest performance among all eight VQA metrics in terms of both PLCC and SROCC values. Thus, in the following section, we will consider VMAF values as reference scores (ground truth) to estimate the quality of the remaining resolution-bitrate pairs which were not assessed during the subjective tests. We then evaluate the rest of the seven VQA metrics on the full dataset (24 reference video sequences and a total of 24 resolution-bitrate pairs, resulting in a total of 576 encoded video sequences). Table 5 shows the PLCC and SROCC correlation values for the seven VQA metrics with VMAF scores. It can be observed that PSNR results in the highest correlation followed by SSIM. Similar to the correlation values with MOS as reported in Table 4 , both RR metrics result in similar performance. Also, it is observed that similar to results reported in Table 4 , for some metrics the correlation values vary significantly over different resolutions. At 1080p, PSNR results in the highest PLCC

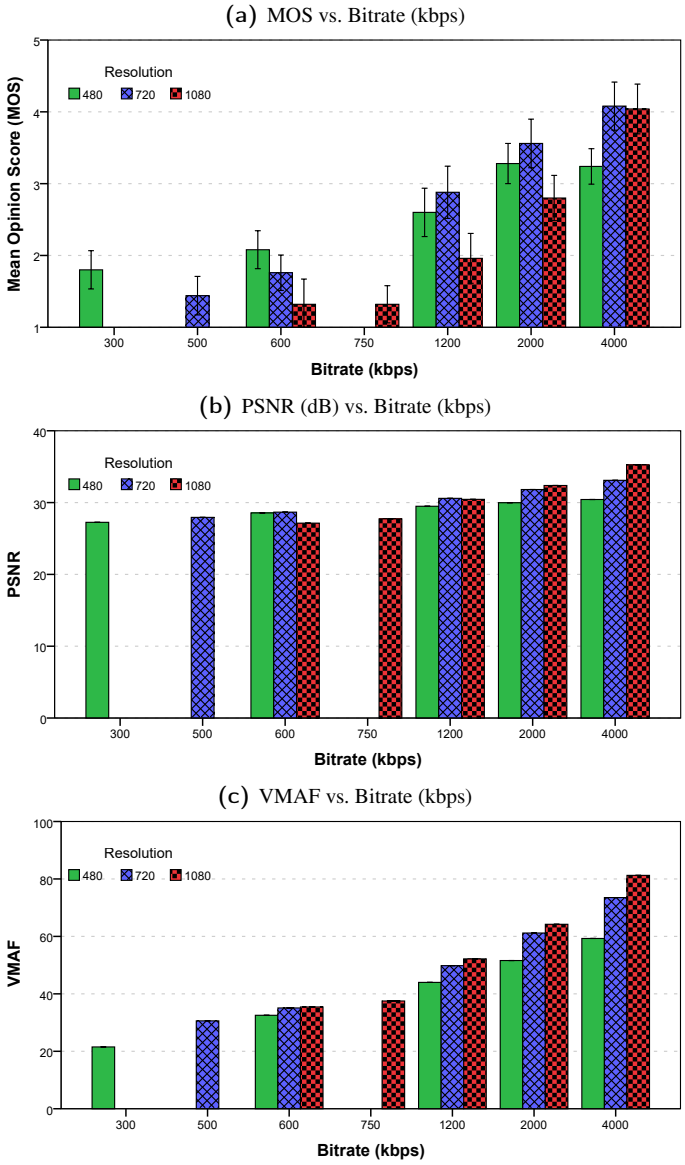

FIGURE 6 MOS (with 95% confidence interval), PSNR and VMAF values for the CSGO video sequence at different resolution-bitrate pairs. A similar behavior is observed for other video sequences (relevant results not reported here due to lack of space).

scores and SpEEDQA results in higher SROCC values. At 720p, SpeedQA results in the highest PLCC and SROCC correlation values. At 480p, NIQE results in the highest PLCC scores and SpEEDQA results in the highest SROCC values. These results indicate towards the high potential for the use of RR and NR metrics for quality evaluations for applications limited to a single resolution and where full reference information is not available.

4.2.4 | Comparative performance analysis of NR metrics

While the VQA metrics, in general, perform quite well, when considering multiple resolutions their performance decreases. Compared to FR and RR metrics, the performance degradation of NR metrics for all data was considerably high. We investigate the reason behind such performance degradation across multiple resolution-bitrate pairs using Figure 7 which shows the scatter plot of BRISQUE, BIQI and NIQE with VMAF scores considering all three resolutions. It can be observed from Figure 7 that, when considering individual resolutions, the variation of the NR metric values with respect to VMAF values are somewhat well correlated and increases linearly and hence results in reasonable PLCC scores. When considering all resolution-bitrate

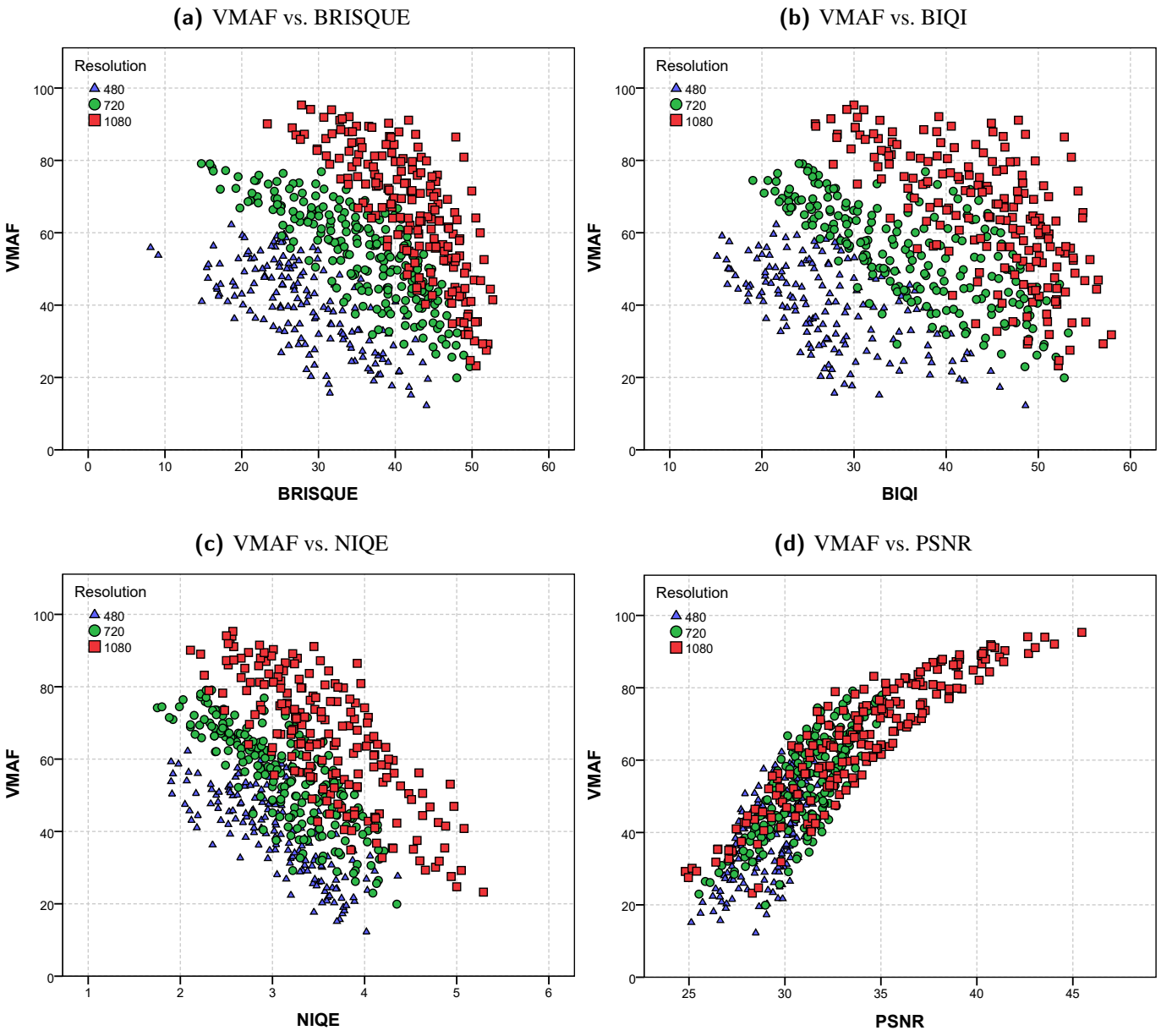

FIGURE 7 Scatter plot showing the variation of the NR metrics and PSNR wrt. VMAF scores considering all three resolutions over the whole dataset.

pairs, however, the spread of values is no longer linear, hence the lower correlation scores. PSNR on the other hand still has a linear correlation when considering all resolution-bitrate pairs and thus results in higher correlation scores. Among the three NR metrics, NIQE results in a much smaller spread for each resolution and when considering all data as compared to BIQI and BRISQUE. Hence, NIQE results in a higher overall prediction quality when using both MOS scores and VMAF scores as the benchmark. BRISQUE, on the other hand, results in almost similar performance as NIQE for 1080p and 720p resolutions but the correlation values decrease for 480p (a wider spread of the scores) and all data. BIQI performs the worst among all three.

The difference in values per resolution can be attributed to the fact that, while for FR and RR metric calculations we used the rescaled YUV videos, for 720p and 480p resolutions, for NR metric calculations we used the downscaled, compressed videos. This, along with lack of proper training with videos consisting of different resolutions, as well as the absence of parameters in the models which can capture the differences due to change in resolution results in lower correlation scores when considering all resolution-bitrate pairs.

TABLE 6 Comparison of the performance of the VQA metric scores with MOS scores in terms of PLCC values for various pooling strategies. The pooling strategy with the highest score for a given VQA metric is shown in bold.

Pooling Method	Objective metrics															
	PSNR		SSIM		VMAF		STRREDOpt		SpeedQA		BIQI		BRISQUE		NIQE	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
Simple Mean	0.74	0.74	0.80	0.80	0.87	0.87	-0.75	-0.77	-0.74	-0.77	-0.42	-0.45	-0.44	-0.46	-0.72	-0.71
I Frame Mean	0.64	0.61	0.68	0.70	0.79	0.79	**-0.79**	**-0.80**	**-0.76**	**-0.80**	-0.46	**-0.48**	**-0.47**	**-0.50**	-0.69	-0.69
Mean Last Frames	0.52	0.54	0.76	0.78	0.78	0.77	-0.58	-0.66	-0.51	-0.66	-0.43	-0.47	-0.45	-0.52	-0.69	-0.70
Minkowski Summation	0.73	0.73	**0.80**	**0.80**	0.86	0.85	-0.74	-0.76	-0.73	-0.77	-0.41	-0.44	-0.44	-0.47	**-0.72**	**-0.71**
Minkowski Exponential Summation	0.53	0.53	0.75	0.77	0.79	0.79	-0.57	-0.66	-0.50	-0.65	-0.45	-0.47	-0.45	-0.51	-0.69	-0.69
P Percentile Lowest_25	**0.81**	**0.81**	0.78	0.79	**0.92**	**0.91**	-0.77	-0.76	-0.76	-0.76	-0.47	-0.45	-0.42	-0.45	-0.65	-0.66
P Percentile Lowest_10	0.79	0.78	0.77	0.78	0.91	0.90	-0.76	-0.75	-0.76	-0.74	**-0.47**	-0.47	-0.39	-0.43	-0.62	-0.63
N Successive Frames	0.77	0.75	0.74	0.78	0.87	0.87	-0.75	-0.73	-0.71	-0.71	-0.43	-0.42	-0.35	-0.38	-0.58	-0.61

4.3 | Effect of Temporal Pooling on VQA performance

The VQA metrics considered in this work provide a score for each frame of the video as most of them like SSIM, NIQE, BRISQUE etc. were initially designed for image quality assessment. For VQA, the usual practice is to consider the average of the individual per-frame level quality scores as the final score of the respective quality metric. Over the years, many different pooling methods such as *Minkowski Summation*, *Mean Last Frames* etc. have been proposed, which are shown to improve the correlation scores for various metrics when compared to simple averaging. An evaluation of six pooling methodologies was initially carried out by Drjle *et al.*[27]. It was found that a proper choice of pooling strategies can significantly improve the prediction quality of a VQA metric. Based on the performance results of various pooling strategies, the authors concluded that taking into account the recency effect and emphasizing the higher importance to low quality segments for the pooling strategies can increase the performance of the VQA metrics. However, this work was limited to six pooling strategies and considered videos of only 352x288 resolution of 12 seconds duration. Therefore, Seufert *et al.*[28] evaluated thirteen pooling methods on long duration videos (100 seconds) considering adaptive streaming application scenarios. Based on the results obtained, it was concluded that none of the complex pooling strategies performed significantly better than that of simple averaging.

Towards this end, we evaluated various temporal pooling strategies considering gaming video streaming application scenario in the present work, which are described in the following. In addition to averaging the individual per-frame level quality scores (*simple mean*), a total of seven different temporal pooling methods were evaluated. The *I-frame mean* pooling method considers the average of all I-frame scores only. In *Mean last frames*, the mean value of the N-last frames is considered (here N=10). The *Minkowski Summation* pooling is defined as $\left[\frac{1}{T}\sum_{i=1}^{T} x^p(t)\right]^{\frac{1}{p}}$, where $x(t)$ is the frame level VQA metric value and p emphasizes the influence of the highest quality frames (here p=2). To take into account the recency effect, the Minkowski Exponential Summation is defined as $\left[\frac{1}{T}\sum_{i=1}^{T} \exp^{\frac{t-T}{\tau}} x^p(t)\right]^{\frac{1}{p}}$, where additionally the weighting factor related parameter τ (here, $\tau = 2$) is introduced. The pooling method *N Successive Frames* calculates the minimum value for mean values of N successive frames (here $N = 10$). Last but not least, the strategies *P Percentile Lowest_10* and *P Percentile Lowest_25* correspond to the mean value of the P percentile lowest quality frames with $P = 10\%$ and $P = 25\%$, respectively. A summary of the performance of the applied pooling methods is shown in Table 6 .

Similar to the results reported by Seufert *et al.*[28], it can be observed that alternative pooling strategies can increase the performance of VQA metrics. However, we could not find a particular pooling method which improves the performance of all VQA metrics or a majority of them. An interesting observation is that, for RR and NR metrics, on an average, the *I-frame mean* pooling strategy results in a higher correlation with subjective ratings as compared to *simple mean* pooling strategy. Considering that in this work an I-frame occurs every two seconds, using the I-frame mean pooling can significantly reduce the computational complexity while resulting in similar performances compared to the other strategies, as the metric evaluation needs to be done only once per two seconds.

TABLE 7 Comparison of the performance of the VQA metric scores with MOS scores in terms of PLCC values for various pooling strategies considering different complexity classes. The pooling strategy with the highest score for a given VQA metric for a certain complexity class is shown in bold.

VQA Metrics	Game Complexity Class	Pooling Strategy								
		Simple Mean	I Frame Mean	Mean Last Frames	Minkowski Summation	Minkowski Exponential Summation	P Percentile Lowest_25	P Percentile Lowest_10	N Successive Frames	TI Pooling
PSNR	High	0.58	0.49	0.25	0.56	0.24	**0.73**	0.67	0.63	0.61
	Medium	0.82	0.62	0.87	0.81	0.87	0.93	**0.94**	0.93	0.86
	Low	0.85	0.78	0.88	0.85	0.88	0.89	**0.89**	0.89	0.85
SSIM	High	0.89	0.72	0.67	**0.89**	0.66	0.85	0.85	0.87	0.86
	Medium	0.77	0.62	0.93	0.77	**0.92**	0.79	0.78	0.73	0.79
	Low	0.79	0.76	0.79	0.79	0.79	**0.80**	0.79	0.79	0.79
VMAF	High	0.81	0.76	0.59	0.78	0.62	0.93	**0.93**	0.88	0.85
	Medium	0.87	0.74	0.89	0.86	0.88	0.92	**0.93**	0.93	0.88
	Low	0.92	0.88	0.91	0.92	**0.92**	0.91	0.90	0.89	0.92
STRREDOpt	High	-0.84	-0.86	-0.52	**-0.87**	-0.51	-0.78	-0.77	-0.86	-0.84
	Medium	-0.86	**-0.89**	-0.77	-0.83	-0.78	-0.88	-0.89	-0.88	-0.84
	Low	-0.64	**-0.65**	-0.60	-0.64	-0.60	-0.64	-0.64	-0.63	-0.64
SpEEDQA	High	-0.81	-0.82	-0.48	-0.83	-0.47	-0.76	-0.77	**-0.86**	-0.81
	Medium	**-0.88**	-0.81	-0.80	-0.86	-0.78	-0.87	**-0.88**	-0.87	-0.86
	Low	-0.63	**-0.66**	-0.60	-0.63	-0.60	-0.63	-0.63	-0.62	-0.63
BRISQUE	High	-0.67	-0.69	-0.63	-0.67	-0.65	**-0.71**	-0.71	-0.68	-0.68
	Medium	-0.45	-0.48	-0.54	-0.45	**-0.56**	-0.47	-0.47	-0.46	-0.45
	Low	-0.05	**-0.08**	-0.07	-0.05	-0.07	-0.04	-0.05	-0.04	-0.05
BIQI	High	-0.72	-0.78	-0.55	-0.71	-0.58	-0.82	**-0.85**	-0.81	-0.74
	Medium	-0.34	-0.42	-0.42	-0.33	-0.49	-0.44	-0.49	**-0.53**	-0.35
	Low	-0.05	-0.08	**-0.11**	-0.05	-0.09	-0.07	-0.07	-0.08	-0.06
NIQE	High	-0.83	-0.83	-0.77	-0.83	-0.78	-0.83	-0.83	**-0.86**	-0.83
	Meduim	-0.56	**-0.59**	-0.59	-0.56	-0.59	-0.51	-0.49	-0.48	-0.56
	Low	-0.76	-0.67	-0.74	-0.76	-0.73	-0.75	-0.75	-0.76	**-0.77**

4.4 | Effect of content complexity on VQA performance

The complexity of a video sequence has a significant effect on the efficiency of video compression. Therefore, for subjective video quality assessment, ITU recommends selecting video sequences that cover a wide range of spatial and temporal complexity [10] as it was done in the present work (see Section 2.2). Consequently, the video complexity may affect the performance of video quality metrics. In this section, we analyze the impact of video complexity on the performance of the metrics that we used. As discussed in Section 3.1), for subjective assessment, we considered six video sequences from three complexity classes: high (CSGO and H1Z1), medium (FIFA18, PC) and low (Dota2, HS). Based on this classification, we evaluated the performance of quality metrics for different content complexity classes and pooling methods, which are summarized in Table 7 . In addition to the pooling strategies discussed in Section 4.3, we used here another temporal pooling strategy *TI pooling*, where the frame level VQA scores are averaged after weighing them using TI scores.

Based on the result in Table 7 , for FR metrics we observe a higher performance for PSNR and VMAF when the video complexity is low, while SSIM performs better for the high video complexity class. One potential reason behind the poor performance of PSNR and VMAF for more complex videos could be the human visual system (HVS) characteristic called visual masking [29]. Visual masking plays a vital role on the perception of distortions in videos as explained by Choi *et al.* [30] who argue that "*more presence of spatial, temporal, or spatiotemporal distortions does not imply a corresponding degree of perceptual quality degradation, since the visibility of distortions can be strongly reduced or completely removed by visual masking*". Choi *et al.* [30] analyzed the motion influences on the performance of VQA metrics by dividing LIVE VQA database [31] into two subsets of small motions and large motions. Their results revealed that some metrics such as PSNR perform poor in case of large motions. However, they did not evaluate VMAF and SSIM. For RR metrics, medium complexity classes result in highest correlation values followed by the high and then low complexity classes.

For NR metrics, we observe that the performance of metrics severely decreases when moving from highly complex content to low complex content, especially for BRISQUE and BIQI. The correlation of BIQI and BRISQUE with subjective video quality ratings for the low complexity class is very low (PLCC of 0.05 for the simple mean method). Although NIQE also performs worse for low and medium complex clusters, the decrease is not as significant as for BIQI or BRISQUE. A possible explanation for the weak performance of NR metrics in the low complex class could be the usage of Natural Scene Statistic (NSS) features in all these three NR metrics which needs to be investigated further. Finally, with respect to the effect of temporal pooling, similar to the results discussed in Section 4.3, there does not exist any particular temporal pooling method which results in the highest correlation value across all metrics even when considering different complexity classes. Also, for RR and NR metrics, similar to earlier, *I-frame mean* pooling seems to perform almost identical to that of *simple mean* pooling across different complexity classes.

5 | CONCLUSION AND FUTURE WORK

Gaming video streaming is an emerging application and is gaining much popularity. We presented in this paper an objective evaluation and analysis of the performance of eight different VQA metrics on gaming video considering a passive, live streaming scenario. Towards this end, a dataset consisting of 24 reference gaming videos of 30 seconds duration and 576 distorted sequences was created. Subjective tests were carried out on a subset of the dataset consisting of 90 video sequences. The performance of the VQA metrics was evaluated on the subjective dataset in terms of PLCC and SROCC values. It was found that VMAF results in the highest correlation with subjective scores. Both RR metrics performance was found to be similar. Among the NR metrics considered in this work, NIQE performed the best. It was observed that many metrics failed to capture the MOS variation at lower resolutions, hence resulting in lower correlation values. Then we evaluated the performance of the rest of the VQA metrics against VMAF on the full test dataset. The performance of the NR metrics decreased when considering different resolution-bitrate pairs together.

Additionally, we analyzed the effect of various temporal pooling strategies on the performance of the VQA metrics. It was found that while for each metric there exists a pooling strategy which results in a higher correlation score than that of commonly used simple averaging, we did not find any pooling strategy which performed significantly better than that of simple averaging across a wide range of metrics. It was found that for RR and NR metrics, I-frame mean pooling strategy resulted in a similar or even better prediction quality as compared to simple averaging, which when used can significantly reduce the complexity as such scores needs to be calculated only a few times over the length of the video, that too only at fixed intervals.

Lastly, we study the performance of the VQA metrics across different complexity classes. It was observed that for VMAF and PSNR, the prediction quality increases while moving across high to low complexity classes while a reverse trend is observed for the NR metrics. SSIM, on the other hand, has a different behavior that the highest correlation appears in the high complexity class followed by low and then medium complexity classes. For RR metrics, medium complexity class performed the best followed by high and low complexity classes. While possible reasons behind this have been discussed in the paper, a further investigation is required.

We believe that the observations and discussions presented in this work will be helpful to improve the prediction efficiency of these existing metrics as well as develop better performing VQA metrics, especially NR metrics with a focus on live gaming video streaming applications. In addition to passive gaming service as discussed in this work, a well-performing NR metric can also be used for predicting video quality for interactive cloud gaming services. It should be noted that our current subjective evaluation was limited in terms of the number of videos considered. Also, the gaming videos used in this work were limited to 30 fps frame rate. As a future work, we plan to extend our subjective analysis using more videos and also include higher frame rate videos and also include more VQA metrics.

ACKNOWLEDGMENT

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 643072 and was supported by the German Research Foundation (DFG) within project MO 1038/21-1.

References

1. Shirmohammadi S, Abdallah M, Ahmed DT, Lu Y, Snyatkov A. Introduction to the special section on visual computing in the cloud: Cloud gaming and virtualization. *IEEE Transactions on Circuits and Systems for Video Technology.* 2015;25(12):1955–1959.
2. Fitzgerald D, Wakabayashi D. Apple Quietly Builds New Networks https://www.wsj.com/articles/apple-quietly-builds-new-networks-13914741492014.
3. Barman N, Zadtootaghaj S, Martini MG, Möller S, Lee S. A Comparative Quality Assessment Study for Gaming and Non-Gaming Videos. In: Tenth International Conference on Quality of Multimedia Experience (QoMEX); 2018; Sardinia, Italy.
4. Barman N, Schmidt S, Zadtootaghaj S, Martini MG, Möller S. An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming. In: Proceedings of the 23rd Packet Video Workshop:7–12; 2018; Amsterdam, Netherlands.
5. Barman N, Zadtootaghaj S, Schmidt S, Martini MG, Möller S. GamingVideoSET: A Dataset for Gaming Video Streaming Applications. In: 16th Annual Workshop on Network and Systems Support for Games (NetGames); 2018; Amsterdam, Netherlands.
6. Fraps . Fraps: Real-time Video Capture and Benchmarking http://www.fraps.com/2018.
7. Claypool M. Motion and scene complexity for streaming video games. In: Proceedings of the 4th International Conference on Foundations of Digital Games:34-41ACM; 2009; Orlando, Florida, USA.
8. Schmidt S, Zadtootaghaj S, Möller S. A Comparison of Interactive and Passive Quality Assessment for Gaming Research. In: Tenth International Conference on Quality of Multimedia Experience (QoMEX); 2018; Sardinia, Italy.
9. Möller S, Schmidt S, Zadtootaghaj S. New ITU-T Standards for Gaming QoE Evaluation and Management. In: Tenth International Conference on Quality of Multimedia Experience (QoMEX); 2018; Sardinia, Italy.
10. ITU-T Rec. P.910 . *Subjective video quality assessment methods for multimedia applications.* 2008.
11. Zadtootaghaj S, Schmidt S, Barman N, Möller S, Martini MG. A Classification of Video Games based on Game Characteristics linked to Video Coding Complexity. In: 16th Annual Workshop on Network and Systems Support for Games (NetGames); 2018; Amsterdam, Netherlands.
12. Twitch . Broadcast Requirements https://help.twitch.tv/customer/en/portal/articles/1253460-broadcast-requirements2017.
13. Barman N, Martini MG. H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 codec comparison for live gaming video streaming. In: Ninth International Conference on Quality of Multimedia Experience (QoMEX):1-6; 2017; Erfurt, Germany.
14. ITU-R Rec. BT.500-13 . *Methodology for the subjective assessment of the quality of television pictures.* 2012.
15. VideoLAN . VLC media player https://www.videolan.org/vlc/2018.
16. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing.* 2004;13(4):600–612.
17. Netflix . VMAF - Video Multi-Method Assessment Fusion https://github.com/Netflix/vmaf2018.
18. Soundararajan R, Bovik AC. Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing. *IEEE Transactions on Circuits and Systems for Video Technology.* 2013;23(4):684-694.
19. Bovik AC, Soundararajan R, Bampis C. On the Robust Performance of the ST-RRED Video Quality Predictor http://live.ece.utexas.edu/research/Quality/ST-RRED/.
20. Bampis CG, Gupta P, Soudararajan R, Bovik AC. Source code for optimized Spatio-Temporal Reduced Reference Entropy Differencing Video Quality Prediction Model http://live.ece.utexas.edu/research/Quality/STRRED_opt_demo.zip2017.

21. Bampis CG, Gupta P, Soundararajan R, Bovik AC. SpEED-QA: Spatial Efficient Entropic Differencing for Image and Video Quality. *IEEE Signal Processing Letters.* 2017;24(9):1333-1337.

22. Mittal A, Moorthy AK, Bovik AC. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing.* 2012;21(12):4695-4708.

23. Moorthy AK, Bovik AC. A Two-Step Framework for Constructing Blind Image Quality Indices. *IEEE Signal Processing Letters.* 2010;17(5):513-516.

24. Mittal A, Soundararajan R, Bovik AC. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters.* 2013;20(3):209-212.

25. Multimedia Signal Processing Group (MMSPG, EPFL) . VQMT: Video Quality Measurement Tool https://mmspg.epfl.ch/vqmt2018.

26. MathWorks . Image Quality Metrics https://www.mathworks.com/help/images/image-quality-metrics.html2018.

27. Rimac-Drlje S, Vranjes M, Zagar D. Influence of Temporal Pooling Method on the Objective Video Quality Evaluation. In: IEEE International Symposium on Broadband Multimedia Systems and Broadcasting:1-5; 2009; Bilbao, Spain.

28. Seufert M, Slanina M, Egger S, Kottkamp M. To Pool or not to Pool: A Comparison of Temporal Pooling Methods for HTTP Adaptive Video Streaming. In: Fifth International Workshop on Quality of Multimedia Experience (QoMEX):52-57; 2013; Klagenfurt, Austria.

29. Sperling G. Temporal and Spatial Visual Masking. I. Masking by Impulse Flashes. *Journal of the Optical Society of America.* 1965;55(5):541–559.

30. Choi LK, Bovik AC. Video quality assessment accounting for temporal visual masking of local flicker. *Signal Processing: Image Communication.* 2018;67:182 - 198.

31. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK. Study of Subjective and Objective Quality Assessment of Video. *IEEE transactions on image processing.* 2010;19(6):1427–1441.

AUTHOR BIOGRAPHY

Nabajeet Barman completed his Bachelor of Technology degree in Electronics Engineering from National Institute of Technology, Surat, India with a focus on Wireless Networks. He then completed MSc. in Information Technology with specialization in Communication Engineering and Media Technology from Universität Stuttgart, Germany during which he worked at Bell Labs, Stuttgart, Germany as part of his internship and master's thesis. Nabajeet is currently a research associate in the Wireless Multimedia and Networking Research Group (WMN) at Kingston University where he is working on QoE-aware video coding strategies as part of MSCA ITN QoE-Net and pursuing his PhD in the field of Quality of Experience of gaming video streaming applications. He is currently Video Quality Expert Group (VQEG) Board member as part of Computer Graphics Imagery (CGI) project and is also involved in ITU-T standardization activities. His research interests include wireless networking, multimedia communications and machine learning.

Saman Zadtootaghaj is a researcher at the Telekom Innovation Laboratories of Deutsche Telekom AG. Since January 2016, Saman is working as a research scientist at the Quality and usability laboratories of Technische Universität Berlin where he is working on Quality of Experience linked to mobile gaming under supervision of Prof. Dr.-Ing. Sebastian Möller. In addition, he is the chair of Computer Generated Imagery group at Video Quality Expert Group (VQEG) and is an active contributor in gaming related work items of ITU-T study group 12.

Steven Schmidt received his Master degree in Electrical Engineering at the Technische Universität Berlin with a major in Communication Systems. Since 2016 he is employed as a research assistant at the Quality and Usability Lab where he is working towards a PhD in the field of Quality of Experience in Cloud Gaming. His research interests include assessment methods for gaming QoE, the identification of influencing factors as well as developing models to predict gaming QoE of cloud gaming services. Therefore, he is involved in gaming-related activities of the ITU Telecommunication Standardization Sector (ITU-T).

Maria G. Martini is a Professor in the Faculty of Science, Engineering and Computing at Kingston University, London, where she also leads the Wireless Multimedia Networking Research Group. She received the Laurea in electronic engineering (summa cum laude) from the University of Perugia (Italy) in 1998 and the Ph.D. in Electronics and Computer Science from the University of Bologna (Italy) in 2002. She has led the KU team in a number of national and international research projects, funded by the European Commission (e.g., OPTIMIX, CONCERTO, QoE-NET, Qualinet), UK research councils, UK Technology Strategy Board/InnovateUK, and international industries. An IEEE Senior Member (since 2007) and Associate Editor for IEEE Transactions on Multimedia (2014-2018), she has also been lead guest editor for the IEEE JSAC special issue and guest editor for the IEEE Journal of Biomedical and Health Informatics, IEEE Multimedia, and the Int. Journal on Telemedicine and Applications, and is currently associate editor for the IEEE Signal Processing Magazine among others. She chaired/ organized a number of conferences and workshops. She is part of international committees and expert groups, including the NetWorld2020 European Technology Platform expert advisory group, the Video Quality Expert Group (VQEG) and the IEEE Multimedia Communications technical committee. She is Expert Evaluator for the European Commission and EPSRC among others.

Sebastian Möller received a Doctor-of-Engineering degree at Ruhr-Universität Bochum in 1999, and his Habilitation in 2004 with a book on the quality of telephone-based spoken dialogue systems. He has been appointed professor at TUB in 2007, is currently Dean of the Faculty IV of EE and CS, Head of Language Technology Lab in Berlin of DFKI, and acts as a Rapporteur for the International Telecommunication Union, ITU-T. He coordinated many activities in the QoE and UX domain, such as WG1 of the COST Action IC 1003 Qualinet (responsible for the Qualinet White Paper on QoE) and is currently the Editor-in-Chief for Springer's Quality and User Experience journal. Sebastian Möller was awarded the GEERS prize in 1998 for his interdisciplinary work on the analysis of infant cries for early hearing-impairment detection, the ITG prize of the German Association for Electrical, Electronic and Information Technologies (VDE) in 2001, the Lothar-Cremer prize of the German Acoustical Association (DEGA) in 2003, a Heisenberg fellowship of the German Research Foundation (DFG) in 2005, and the Johann Philipp Reis prize in 2009. Since 1997, he has taken part in the standardisation activities of the International Telecommunication Union (ITU-T) on transmission performance of telephone networks and terminals.

How to cite this article: N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller (2018), An Objective and Subjective Quality Assessment Study of Passive Gaming Video Streaming, *International Journal of Network Management*, *2018;00:1–6.*