QUALITY-OF-EXPERIENCE OF STREAMING VIDEO: INTERACTIONS BETWEEN PRESENTATION QUALITY AND PLAYBACK STALLING

Kai Zeng, Hojatollah Yeganeh and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada

ABSTRACT

Network streaming video services have been growing explosively in the past decade, but how to measure and assure the video quality-of-experience (QoE) of end consumers is still an open problem. Poor presentation quality and playback stalling have been identified as the most dominant factors that degrade user QoE. Although both factors have been studied individually, little is known about the interactions between them. In this work, we first construct a streaming video database that contains compressed videos at different distortion levels and with different stalling patterns. We then carry out a subjective test to evaluate the QoE of the videos. The results reveal some interesting dependency between presentation quality and playback stalling. Specifically, playback stalling always causes QoE degradation, but the strength of such degradation depends on the presentation quality when the stalling event occurs.

Index Terms— Video Stalling, Presentation Quality, Video Streaming, Quality-of-Experience

1. INTRODUCTION

The explosive growth of streaming media services in the past decade is creating many technical challenges, among which how to assure the visual quality-of-experience (QoE) of end consumers is one of the most critical ones. A recent large scale survey [2] shows that 75% of the consumers will switch their video channels if they experience a subpar service for more than 5 minutes. Youtube launched its Video Quality Report program [3] to help its customers from worldwide locations to understand the performance of their Internet Service Provider (ISP) and other factors that may effect the viewing experience. Netflix releases a country-based ISP speed index [4] to help its customers to gain more knowledge about the expected video QoE. A common major problem of such real systems is the lack of a direct way to quantify the real QoE of end users. Rather, indirect parameters such as video bitrate have been used to give a rough estimate of visual QoE. But bitrate could be an extremely poor indicator of the actual video quality. Therefore, an accurate, real-time, and easy-to-use video QoE measurement tool is highly desirable.

Stalling is considered as a highly annoying quality issue, which could cause strong user frustration during a viewing session [3]. Based on the 2015 Conviva viewer experience report, 28.8% of viewing experience suffers from stalling problem, which causes significant reduction of viewer engagement. Significant effort has been made to understand the perceptual impact of stalling and discuss the way to reduce stalling. Hossfeld et al. [5] demonstrated that subjects are extremely sensitive to stalling during playback, and service provides should design their systems accordingly, for example, by increasing the initial loading time or the compression ratio. Ghadiyaram et al. [6] conducted a large-scale subjective experiment to study the stalling events on the QoE of mobile streaming videos, and analyzed the impact of the frequency, position and length of stalling occurrences. Garcia et al. [7] focused on the progressive download type of video services and investigated the quality impact due to initial loading, stalling, and compression for HD sequences. They observed an additive impact of stalling and compression on the perceived QoE and reported that the perceptual impact of stalling is independent of the video content at high bitrate. Staelens et al. [9] evaluated the video stalling effect due to camera feed switch by a subjective study, towards the overall quality ratings for adaptive streaming of sports event.

Another major QoE factor is the presentation quality of the compressed video (without considering transmission issues such as rebuffering and switching). Traditional objective video quality assessment (VQA) models, such as SSIM and MS-SSIM, are useful in measuring the presentation quality, but lack certain features that are important in video streaming scenarios. A QoE database for HTTP-based video streaming was constructed that contains compressed sequences with temporally variable bitrates to study time-varying subjective quality of rate-adaptive videos [12]. Aiming at IP-based video streaming, the MCLV database was built, which contains videos with 4 compression and 2 spatial scaling levels [13]. State-of-the-art VQA metrics had been shown to have moderate prediction performance on this database. Lievens et al. [14] reported that the classical quality measurements (PSNR, SSIM, VQM) failed to predict the perceived video quality in HTTP adaptive streaming due to quality fluctuations, and proposed an empirical quality metric to account for the streaming-specific distortions. A device adaptive video QoE measurement, named SSIMplus [10, 11], was designed to extend the capabilities of VQA methods to streaming application. A number of unique features, such as real-time speed, cross-resolution assessment, automatic alignment, device adaptivity, make SSIMplus a much better tool suited for streaming media applications.

A straightforward way to reduce the probability of stalling is to reduce bitrate, but reducing bitrate will meanwhile lower presentation quality. Apparently, achieving an optimal compromise between the two is critical. For this purpose, a reliable QoE measure that considers both factors is important. Unfortunately, very few efforts have been made to investigate the joint effect of them. In [15], it is stated that "if the video stalls, the video experience of the user is disturbed - independent of the actual video characteristics", which means stalling and presentation quality are independent of each other, a similar conclusion as in [7]. In this work, however, this conclusion is challenged by our experimental results. More specifically, we built a new database that contains compressed videos at different distortion levels and with different stalling patterns, and then conducted a subjective test to evaluate the QoE of the videos. Among a number of observations that are useful for the future development of a complete QoE prediction model, an interesting one is that playback stalling has stronger negative impact if the presentation quality is higher when the stalling event occurs.

Fig. 1. Test video sequences.

2. SUBJECTIVE QUALITY ASSESSMENT

To the best of our knowledge, existing relevant databases in the literature focus on either video presentation quality only (where the distortion is mostly compression) or the impact of stalling only (for different patterns of stalling position, duration, and frequency), making it difficult to observe the dependencies between presentation quality and playback stalling. Therefore, our first goal here is to develop a new database that can be used to study the interaction between them.

Twenty high-quality video sequences of 1920x1080 resolution and 10-second [8] long are selected to cover diverse video content, including animation, humans, plants, natural sceneries, movie, sports, live show, indoor and outdoor views, and man-made architectures. The detailed specifications of those videos are listed in Table 1 and a screenshot of each video is shown in Fig. 1.

Using aforementioned sequences as the source, each video is encoded into three bitrate levels, 500Kbps, 1000Kbps, 1500Kbps, to cover different quality levels. A 5-second stalling event is inserted at either the beginning or the middle point of the encoded sequences. In total, we obtain 200 test samples that include 20 source videos, 60 compressed videos, 60 initial stalling videos, and 60 mid-stalling videos. An example of a stalling frame is shown in Fig. 2.

The test video sequences are categorized into four groups: A: reference group; B: videos with compression artifacts only; C: videos with compression and initial stalling; D: videos with compression and middle stalling. Groups A and B can be used to test VQA metrics that aim for predicting presentation quality. Using Groups A and B as the anchors, Groups C and D are useful to study the impact of stalling.

A computer with Intel(R) Core(TM) i7-2600 dual 3.40GHz CPU was used in the subjective user study. The test environment was setup as a normal indoor office workspace with ordinary illumination level. All videos are displayed at their actual pixel resolution on an LCD monitor at a resolution of 2560×1600 pixels with Truecolor (32bit) at 60Hz. The monitor was calibrated in accordance with the recommendations of ITU-T BT.500 [16]. A customized graphical user interface (GUI) was used to render the videos on the screen with random order during the test.

We adopted a single-stimulus quality scoring strategy. A total of 25 naïve observers, including 13 males and 12 females aged between

Table 1. The details of reference videos.

Index	Name	Frame Rate	Bitrate(Mbps)
1	Animation	25	32.7
2	Biking	50	297
3	BirdsOfPrey	30	114
4	ButterFly	25	163
5	CloudSea1	24	163
6	CloudSea2	24	133
7	CostaRica1	25	131
8	CostaRica2	25	128
9	Football1	25	140
10	Football2	25	90.3
11	Football3	25	95.3
12	Forest1	25	281
13	Forest2	25	247
14	MTV	25	259
15	Ski	30	148
16	Squirrel	25	136
17	Transformer1	24	102
18	Transformer2	24	156
19	Basketball1	25	137
20	Basketball2	25	217

Fig. 2. An example of a stalling frame.

22 and 30, participated in the subjective experiment. For each subject, the whole study takes about one and half hour, which is divided into three sessions with two 7-minute breaks in-between. In order to minimize the influence of fatigue effect, the length of a session was limited to 25 minutes. During the test, subjects were asked to watch a single video at one time, and give their opinions about the video QoE after playback. The user interface of the scoring panel is shown in Fig. 3. The score ranges between 0 and 100, where 0 is the worst quality and 100 is the best.

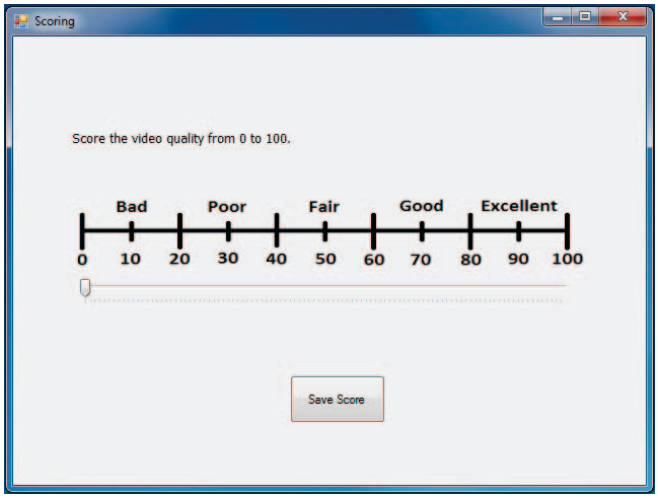

Fig. 3. The scoring panel used in the user study.

3. ANALYSIS

After the subjective user study, 2 outlier subjects are removed based the outlier removal scheme in [16], resulting in 23 valid subjects. All collected subjective scores are converted to Z-scores based on the sample mean and standard deviation. The final quality score for each individual video is computed as the average of the subjective scores, i.e., the mean opinion score (MOS), from all valid subjects.

Considering the MOS as the "ground truth", the performance of an individual subject can be evaluated by comparing his/her quality scores with the "ground truth" for all test images. The Pearson linear correlation coefficient (PLCC) and Spearman's rand-order correlation coefficient (SRCC) are employed as the comparison criteria. Both PLCC and SRCC range between 0 and 1, within which higher values indicate better performance. This is done for each individual subject and the results for all subjects are depicted in Fig. 5. It can be seen that most individual subjects perform reasonably well in terms of predicting MOS scores. The average performance across all individual subjects and the standard deviation between them are also given in the rightmost columns in Fig. 5. This provides a general idea about the performance of an average subject. In conclusion, there is a considerable agreement between different subjects on the quality of the test sequences.

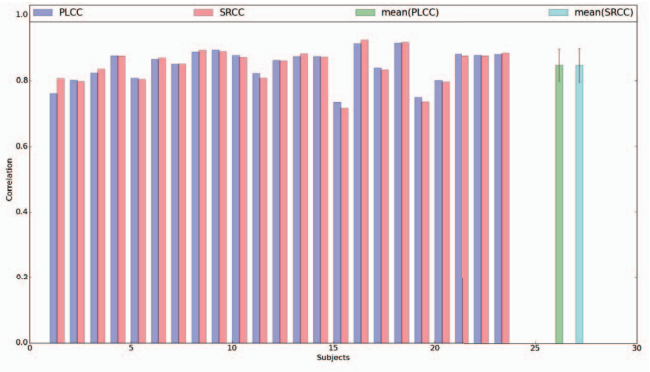

Fig. 5. Performance evaluation of individual subjects using MOS as the ground truth.

Video sequences in Groups A and B do not involve any stalling event, and thus can be used directly to test VQA models designed for predicting presentation quality. The models being tested include both classical and state-of-the-art full-reference (FR), reduced-reference (RR) and no-reference (NR) methods. The test results are given in Table 2, where PLCC and SRCC are used as the performance evaluation criteria. Not surprisingly, in general RR models perform better than NR models, and FR models better than RR models. Among all models, the SSIMplus [10, 11] and MS-SSIM [18] models achieve the best performance, which provide good predictions of the presentation quality. However, since these models do not consider stalling, they are not able to provide any further insight about visual QoE degradation in the presence of initial buffering and rebuffering during video playback.

Table 2. Performance of objective VQA models on test video sequences without stalling.

IQA model	Type	PLCC	SRCC
PSNR	FR	0.7186	0.7127
SSIM[17]	FR	0.7779	0.7650
MS-SSIM[18]	FR	0.8817	0.8412
SSIMplus[10]	FR	**0.8829**	**0.8595**
VQM[19]	FR	0.7497	0.7460
RRIQA[20]	RR	0.7169	0.7083
STRRED[21]	RR	0.8196	0.8076
BIQI[22]	NR	0.5952	0.5624
NIQE[23]	NR	0.5371	0.5524
BRISQUE[24]	NR	0.5120	0.5486

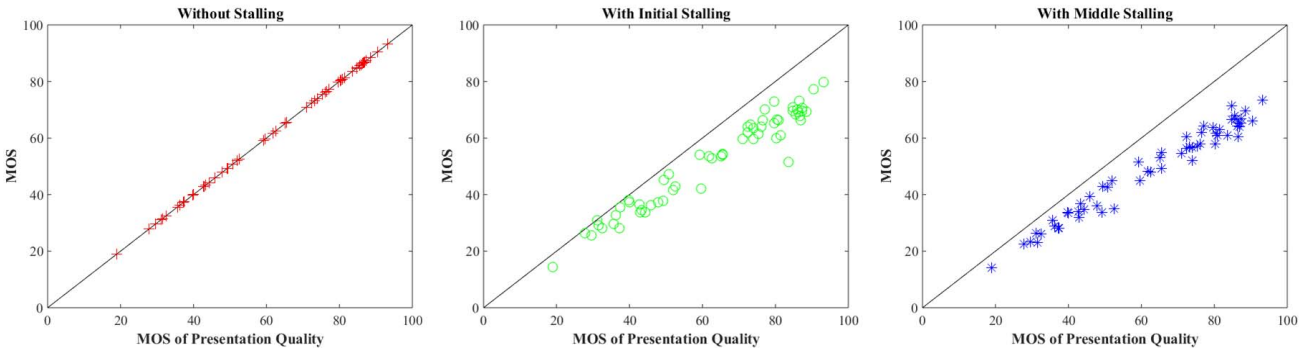

Fig. 4. The impact of stalling on the drop of MOS values.

As discussed earlier, the key question we would like to answer through this study is whether the impact of the stalling events is independent of video presentation quality. If they are independent, then regardless of the presentation quality level, stallings will have the same impact on the overall video QoE scores. Therefore, assuming the additive relationship between stalling and presentation quality is valid as in [7], we are expecting a near constant quality drop when stalling occurs in videos with different presentation quality and compression levels.

Fig. 4 depicts the scatter plots between the MOS values of video presentation quality (MOS values of videos in Group B) versus the MOS values of the corresponding videos without stalling (Group B), with initial stalling (Group C), and with middle-stalling (Group D). There are several useful observations. First, stalling always results in drops in MOS, and the drops could be very significant. Second, the impact of stalling in the middle of the video is in general stronger than that at the beginning of the video. This is verified with the numerical statistics shown in Table 3, where comparing the last two rows, we concludes that the drops in MOS in all three compression levels are consistently higher in the mid-stalling than in the initial stalling cases. A reasonable explanation is that viewers are more tolerant to the stalling at the beginning which is often created by initial loading in practice. By contrast, stalling in the middle creates discontinuity in consuming the video content and is thus more annoying. This is a phenomenon observed in previous studies, and is also verified during our discussions with the subjects after they finished their subjective tests. Third, the drop in MOS is not a constant, but varies with the increase of the MOS of the presentation quality. This can also be seen in the last two rows of Table 3, where in both initial stalling and mid-stalling cases, the MOS drops increases significantly with bit rate levels. This suggests that stalling is creating stronger frustration on the viewers when it occurs in videos with high presentation quality. One explanation may be that users have higher expectations when watching high quality video, thus when stalling happens, it is less expected and results in stronger dissatisfaction.

We find that an empirical model that may be used to describe the MOS drops observed in the current subjective data for both the initial stalling and middle stalling cases (but with different sets of parameters) is given by

$$\Delta Q = c_1 Q + c_2 Q^2 + c_3 Q^3 \,, \tag{1}$$

where Q is the presentation quality without stalling, and ΔQ is the drop of quality caused by stalling, measured by the difference of MOS (DMOS) of the videos with and without stalling. The fitting parameters for the cases of initial and mid-stalling are given by

Table 3. The average MOS values for different stalling events.

Bitrate(Kbps)	500	1500	3000
No Stalling	42.29	66.23	81.67
Initial Stalling	35.99	55.27	66.01
Middle Stalling	33.29	52.42	61.63
NoStalling − InitialStalling	6.30	10.96	15.66
NoStalling − MiddleStalling	9.00	13.81	20.04

$\{-2.5e-6, -0.7e-3, -0.11\}$ and $\{-1.4e-5, 0.001, -0.22\}$, respectively. Fig. 6 shows how the empirical models fit the subjective data.

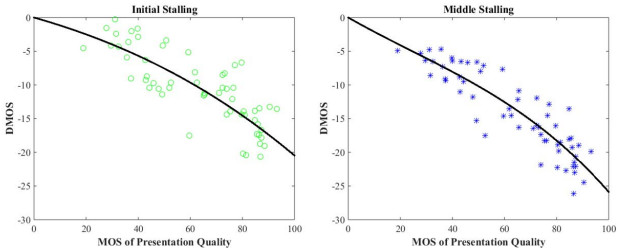

Fig. 6. The impact of stalling along with the change of presentation quality.

4. CONCLUSION

We built a database and conducted a subjective user study to investigate the visual QoE of streaming video. In particular, we make one of the first attempts that focuses on understanding the interactions between presentation quality and playback stalling. The most important finding of the current work is that unlike widely assumed in previous studies, the impact of presentation quality and playback stalling on the overall QoE is not independent. Instead, the negative impact of stalling increases significantly with the level of presentation quality when the stalling occurs. We use a simple model to describe this trend. We believe this work will provide useful insights in the future development of comprehensive QoE models for video streaming.

5. REFERENCES

[1] Cisco, "Cisco visual networking index: global IP traffic forecast, 2014-2019", White Paper, May (2015).

[2] "Conviva's consumer survey report 2015: how consumer judge their viewing experience." http://www.conviva.com/csr-2015/.

[3] "Google Video Quality Report", https://www.google.com/get/videoqualityreport/.

[4] The ISP speed Index from Netflix, http://ispspeedindex.netflix.com/.

[5] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch and C. Lorentzen, "Initial delay vs. interruptions: between the devil and the deep blue sea", Fourth Int. Workshop on Qualty of Multimedia Experience (QoMEX), pp. 1-6, Yarra Valley, VIC, July 5-7, 2012.

[6] D. Ghadiyaram, A.C. Bovik, H. Yeganeh, R. Kordasiewicz and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos", IEEE Global Conf. on Sig. and Info. Proc. (GlobalSIP), pp. 989-993, Atlanta, GA, Dec. 3-5, 2014.

[7] M.N. Garcia, D. Dytko and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services", Sixth Int. Workshop on Quality of Multimedia Experience (QoMEX), pp. 129-134, Singapore, Sept. 18-20, 2014.

[8] P. Frohlich, S. Egger, R. Schatz, M. Muhlegger, K. Masuch, B. Gardlo, "QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment?", Fourth Int. Workshop on Quality of Multimedia Experience (QoMEX), pp.242-247, Yarra Valley, VIC, July 5-7, 2012.

[9] N. Staelens, P. Coppens, N. Van Kets, G. Van Wallendaef, W. Van den Broech, J. De Cock and F. De Turck, "On the impact of video stalling and video quality in the case of camera switching during adaptive streaming of sports content", Seventh Int. Workshop on Quality of Multimedia Experience (QoMEX), pp. 1-6, Pylos-Nestoras, May 26-29, 2015.

[10] A. Rehman, K. Zeng and Z. Wang, "Display device-adapted video quality-of-experience assessment", IS&T-SPIE Electronic Imaging, Human Vision and Electronic Imaging XX, Feb. 2015.

[11] The SSIMplus Index for Video Quality-of-Expeirnce Assessment, http://ece.uwaterloo.ca/ z70wang/research/ssimplus.

[12] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath and A. C. Bovik, "Modeling the time-varying subjective quality of HTTP video streams with rate adaptations", IEEE Trans. Image Processing, vol. 23, no. 5, pp. 2206-2221, May, 2014.

[13] J.Y. Lin, R. Song, C.-H. Wu, TJ Liu, H. Wang and C.-C. J. Kuo, "MCL-V: a streaming video quality asessment database", Journal of Visual Communication and Image Representation, vol. 30, pp. 1-9, July 2015.

[14] J. Lievens, A. Munteanu, D. De Vleeschauwer and W. Van Leekwijck, "Perceptual video quality assessment in HTTP adaptive streaming", IEEE Int. Conf. on Consumer Electronics (ICCE), pp. 72-73, Las Vegas, NV, Jan. 9-12, 2015.

[15] T. Hossfeld, R. Schatz, E. Biersack and L. Plissonneau, "Internet video delivery in Youtube: From traffic measurements to quality of experience", in Springer book on Data Traffic Monitoring and Analysis: From measurement, classification and anomaly detection to Qualty of experience, 2013.

[16] ITU-R BT.500-12, Recommendation: methodology for the subjective assessment of the quality of television pictures, Nov. 1993.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Trans. Image Processing, vol.13, no.4, pp. 600-612, Apr. 2004.

[18] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-scale structural similarity for image quality assessment", IEEE Asilomar Conf. Signals, Systems and Computers, Nov. 2003.

[19] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality", IEEE Trans. Broadcasting, vol. 50, no. 3, pp. 312-322, Sept. 2004.

[20] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model", Human Vision and Electronic Imaging X, Proc. SPIE, vol. 5666, San Jose, CA, Jan. 2005.

[21] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing", IEEE Trans. Circuits and Systems for Video Technology, vol. 23, no. 4, pp. 684-694, Apr. 2013.

[22] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices", IEEE Signal Processing Letters, pp. 513-516, vol. 17, no. 5, May 2010.

[23] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a completely blind image quality Analyzer", IEEE Signal Processing Letters, pp. 209-212, vol. 22, no. 3, March 2013.

[24] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-reference Image Quality Assessment in the spatial domain", IEEE Trans. Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012.