5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)
29-31 August 2016, Berlin, Germany

Enhancing Video Streaming Using Real-Time Gaze Tracking

Sebastian Arndt[1], *Jan-Niklas Antons*[2]

[1]Norwegian University of Science and Technology, Trondheim, Norway
[2]Technische Universität Berlin, Berlin, Germany
sebastian.arndt@ntnu.no, jan-niklas.antons@tu-berlin.de

Abstract

While users are watching videos, they can only focus and evaluate one small part of the presented video frame which is within the focal view point. Video coding however often assumes equal distribution of attention or has predefined areas of interest, regardless of where the observer is actually looking at. In this paper, we propose a system that uses real-time information of eye gaze in order to perform video coding more appropriate to the viewers' attention focus. Using subjective quality evaluation performed after using a prototype, we show that test participants evaluate attention focus based coding better than traditional coding. On the long run this result can stimulate more research in the domain of real-time gaze based video coding.

Index Terms: Quality of Experience, Eye Tracking, Adaptive Streaming, Real-time, Eye Gaze

1. Introduction

When transmitting videos via the internet, a compromise between the transmitted bit rate and the experienced quality at the user's side has to be found. Based on the existing network capacities, the transmitted video bit rate has to be adjusted accordingly. Two techniques are usually applied to compensate for this: introducing stalling events and/or reducing the bit rate of the video stream. In case of stalling, two different techniques can be considered for video transmission, either at the beginning a longer waiting time has to be taken into account, or stalling happens in between while the consumer is watching the video. During stalling the video stream stops and buffers enough material so that it can playback the video smoothly again. The latter case should be avoided as the experienced quality decreases significantly introducing these stalling event, as mentioned in [1]. However, this is hardly possible in case of live events.

When reducing the bit rate, the standard procedure is that the whole video frame has a lower bit rate, and thus, lower quality when the network conditions are not sufficient for the best possible bit rate. Here, MPEG-DASH (dynamic adaptive streaming over HTTP) is the current standard to vary the transmitted bit rate depending on the current network traffic [2]. As the result of bit rate variation, the experienced quality varies as well. In some cases, an optimum between stalling events and a decrease in quality has to be identified. It was shown that both degradation types have similar influences on the perceived quality according to [3].

To test, whether one streaming system leads to better quality of experience than another, subjective quality tests are often performed. Here, participants are watching test sequences under different technical conditions and afterward have to rate these concerning the perceived quality on either an absolute category scale (ACR) or degradation category scale (DCR). Averaging the individual quality ratings over all test participants for one condition, leads to the mean opinion score (MOS) [4].

The area of interest (ROI) of a video describes that part of the video frame where the observer usually looks at. The general idea to put more focus on the ROI in a frame during video playback and to neglect the rest of the video area in terms of bandwidth has been proposed in [5]. Initially ideas were employed in which only the middle of the video was not affected by degradations. In [6], the peripheral area was distorted while the middle was without any degradation. The test participants had a search task to perform. Here, it was shown that when degrading the peripheral area subject's performance was almost not affected.

In [7] and [8], a system has been introduced that uses region of interest-based adaptive streaming (ROIAS). Here, the area of interest was calculated on the basis of a viewer region of interest model. The quality is decreasing either linearly or logarithmically from the center ROI. Using objective and subjective measures, it was shown that ROIAS is beneficial when bandwidth is an issue. Lee et al. provide an extensive review of different perceptual video compression metrics [9].

In scenarios where only one person is watching the video, more individual streaming might be of more interest in case transmitted bandwidth must be reduced. Such that, the current visually attended area is in high quality and the video in the peripheral area may be of lower quality, as the peripheral area does not matter as much to the user in terms of quality evaluation. Therefore, eye trackers can become handy. Especially as the market for low cost eye tracking systems is rising which makes it interesting also for consumers.

In the this paper, we present an approach to vary the bit rate of the video by using information of the observer's gaze. Only within the fovea the human perceives the observed scene sharp. The fovea consists of less than 2° of the eye ball, and the further away from the fovea a scene appears the less detailed it is perceived by the human. This is due to an over-representation of cones in this area, and a much looser concentration of those in the rest of the human eye. Cones are part of the retina and are responsible for color vision in the human. The two areas can be clustered as foveal and peripheral sight [10]. Thus, it is assumed that only roughly 5° of visual angle are of relevance when observing a visual scene [11]. This leads to the assumption that it is not necessary to transmit data for the peripheral area in full bit rate.

The above mentioned approaches are static and do not involve the user who currently is watching the video. Thus, if the user would move his gaze outside the predefined ROI, the experienced quality would decease dramatically. The system implemented for this paper uses real-time eye tracking data and

10.21437/PQS.2016-3

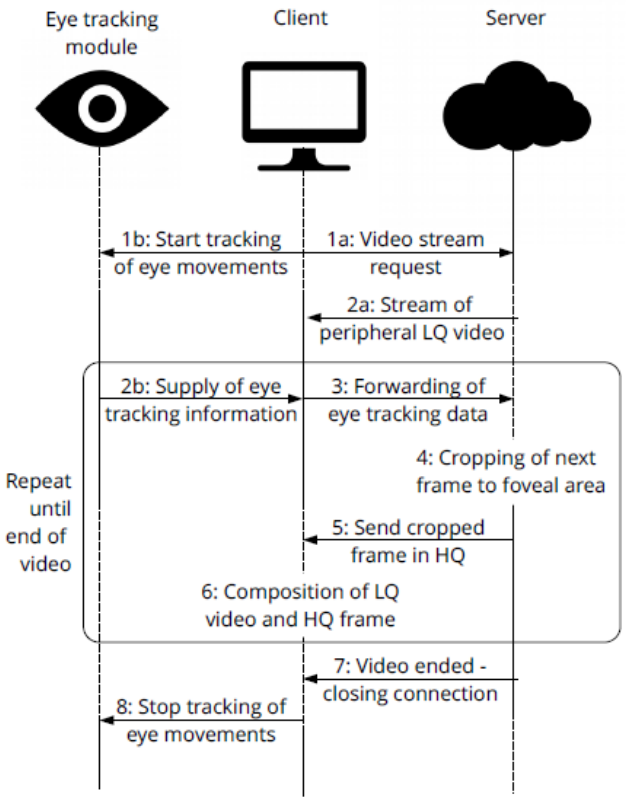


Figure 1: Schematic representation of the technical setup and implemented protocol. Figure is taken from [12].

gives the current region of attention the highest bit rate. As the human foveal area is circular, the calculated areas are circular as well. The research question is whether there exists an ideal configuration of the size of the high-quality area and the bit rate reduction on the surrounding areas.

In Section 2 we describe the technical apparatus and software implemented for this study. Section 3 describes the experimental setup and stimulus used. In Section 4 the results of the experiment are presented. The paper concludes with Section 5 where also future directions are discussed.

2. Technical Setup

The tested setup consisted of a standard notebook (client) which runs the developped application, a virtual server hosting the videos, and an eye tracking device (see Figure 1 for detailed setup of the system).

The eye tracking device used was a low-cost device from The Eye Tribe. This is a remote eye tracking device, placed below the monitor. The device has a sampling frequency of 60 Hz and an accuracy provided by the manufacturer of $0.5^\circ - 1^\circ$.

After starting the application on the client, it requests the video stream from the server *(1a)* and the live data stream from the eye tracker *(1b)*. Regardless of the real-time gaze data, the server already starts to send the video stream *(2a)* in order to minimize initial delay and in case no area is focused. The client receives the gaze data from the eye tracking module *(2b)* and forwards it to the server *(3)*. Whenever a fixation was detected

Figure 2: Example frame of the video. *X* depicting the fixation point of the observer, *F* being the foveal area shown in high quality, and *P* being the peripheral area which is shown in reduced bit rate. Figure is taken from [12].

r / q	0	150	190	230	270
% of frame affected	**100**	**7.67**	**12.31**	**18.03**	**24.85**
20	1	1	1	1	1
26	0.49	0.53	0.56	0.58	0.62
32	0.28	0.34	0.37	0.41	0.46
38	0.18	0.25	0.28	0.33	0.39
44	0.14	0.21	0.25	0.30	0.35
51	0.12	0.18	0.23	0.28	0.34

Table 1: Overview on achieved bit rate reduction in percentage (i.e. 1 equals original bit rate). Rows depict level of quantization *q*, and columns radius size of the foveal circle, *r*, in px.

by the eye tracker the current gaze point is updated. Based on this, the video sent is the processed video containing high-quality at the fixation area and low-quality in the peripheral area *(4)*. The implemented system outputs a single video with two different compression rates *(5, 6)*. This was achieved such that on top of the low-quality peripheral video frame a video in high-quality was overlayed at exactly that area the participant was focusing on. When no area was fixated the video was played back fully in the high-quality version. After the video is finished, the connection to the server is closed *(7)* and the eye tracker stops sending gaze information *(8)*. In order to reduce delay, the cropping is done on the client's side at the moment.

3. Method

3.1. Participants

26 test participants took part in the experiment (21 male, 5 female) with an average age of 25.1 years (ranging from 21 to 41 years). All participants had normal or corrected-to-normal vision.

3.2. Stimulus

The stimulus used was a 15 s excerpt from the freely available animation movie 'Big Buck Bunny'. It showed a scene of three characters and a fourth coming into the picture and mov-

ing around. The video was encoded using the x264 encoder. The area provided in high-quality was encoded using constant quantizer mode of r = 20. The low-quality versions for the peripheral video were played with q = 51, 44, 38, 32, and 26. The gaze-concentric circles had a radius of r = 150, 190, 230, and 270 pixels. Additionally, the whole video was manipulated using the mentioned quantization rates (indicated by $r = 0$). All combinations of circle sizes and compression rates were tested, resulting into 25 conditions. Each condition was evaluated once by each participant. The order of conditions were randomized between participants. The video was shown with 25 fps. In Table 1 approximations for the video reduction rate with each configuration can be seen. It assumes equal bit rate distribution over the entire frame.

Sound was played via loudspeakers positioned left and right in front of the test participant. Audio contained only animated sounds and no speech. The volume was set to a comfortable level by the experimenter and was at the same level for all participants.

3.3. Experimental Design

Participants gave informed consent about the experiment. Following this, they filled in a demographic questionnaire as well as a questionnaire concerning their video consumption behavior and previous experience with quality and eye tracking experiments. After a trainings trial, the actual experiment started which lasted approximately 20 min. Afterward a post-questionnaire was provided where participants could give feedback towards the system and give advice on things they noticed during the experiment.

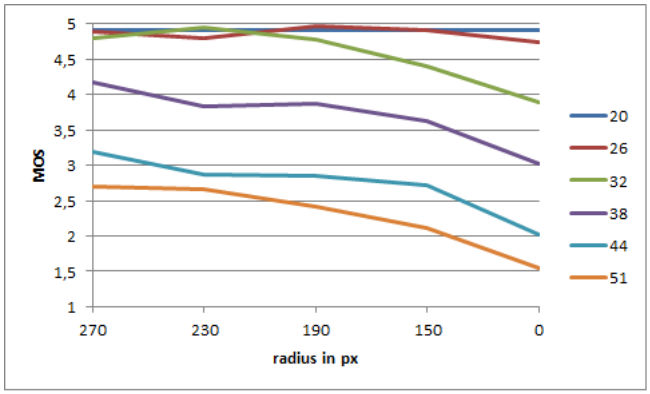
After each trial (i.e. playing of one video) participants were asked for an individual rating of video, audio, and audiovisual quality, on a 7-point continuous ACR scale (as in ITU-T P.911 [13]) with transition areas (i.e. ranging from 'extremely bad' to 'ideal'). The quality judgment was asked for on the screen after playing the stimulus.

4. Results

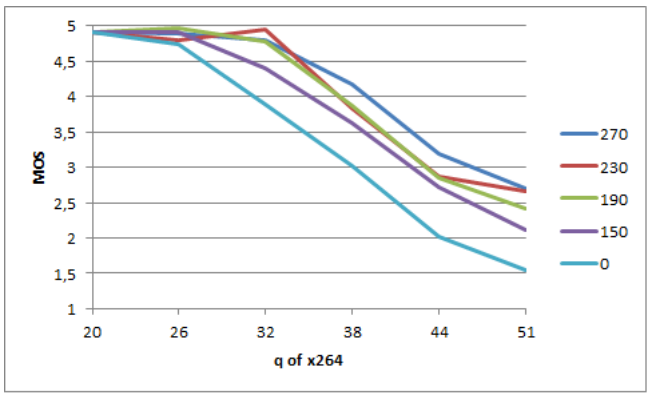
In Figure 3a, it can be seen, how the evaluated quality is affected by the high quality radius size. The MOS ratings are higher for larger radius sizes, calculating a repeated measures ANOVA with MOS as the dependent variable, and radius size (5 levels) as independent variable reveals statistical significance ($F(4, 100) = 31.69$, $p \leq 0.01$). Bonferoni corrected pairwise comparisons reveal that the perceived quality is higher with the smallest circle (r = 150 px) compared to the complete frame in reduced quality. The r = 270 px circle is perceived significantly better compared to r = 150 px and r = 190 px. All other pair-comparisons did not differ significantly in their perceived quality.

In Figure 3b, the dependency of the peripheral compression and perceived quality is shown. The MOS is increasing with decreasing quantization. Statistical significance is obtained when calculating a repeated measures ANOVA with MOS as dependent variable and quantization rate (5 levels) as independent variable ($F(5, 125) = 116.31$, $p \leq 0.01$). Furthermore, the data shows significantly higher ratings for q = 51, 44, and 38 compared to all other conditions.

Distorting the entire frame with each respective quantization rate was evaluated significantly worse compared to the conditions where the focus area was of high quality and only the peripheral area was distorted.



(a) Quality ratings of quantization level relative to the affected radius.



(b) Quality ratings of affected radius relative to the quantization level.

Figure 3: Averaged quality ratings over all participants.

Additionally, there is an interaction effect between radius size and peripheral compression rate ($F(20, 500) = 5.47$, $p \leq 0.01$). Comparing the individual settings with each other, following optimal configurations can be found: r = 230 and q = 32; r = 190 q = 32; r = 150 q = 32; and r = 0 q = 26.

Furthermore, the audio quality and the audiovisual overall quality was rated after each video. Here, no significant effect could be observed for the audio MOS with foveal size as the independent variable ($F(4, 100) = 1.51$, n.s.). In case of the overall audiovisual MOS a significant decline could be observed when the size of the high-quality area was varied ($F(4, 100) = 27.53$, $p \leq 0.01$). With the peripheral video quality as an independent variable a significant effect could be observed for audio MOS ($F(5, 125) = 7.07$, $p \leq 0.01$) as well as for audiovisual MOS ($F(5, 125) = 102.88$, $p \leq 0.01$).

To reveal problems with the implementation after the test a short post-questionnaire was conducted. Here, it was revealed that almost all participants saw a circle of the high-quality focus area; and roughly half the subjects (11 out of 26) noticed a delay between the focus and peripheral area.

5. Conclusion

Testing the implemented system, it was shown that already a small circle in the main focus area of the viewer improves the experienced quality compared to cases where the entire frame had a reduced bit rate. The larger this circle gets the better are the quality ratings. However, the quality of the peripheral area also is an important factor which contributes to the overall quality.

The proposed system shows a significant increase in perceived quality, when only the area which is not focused by the observer is in reduced quality compared to when the whole frame is affected by quality reduction.

In this test, both the foveal size and bit rate of the peripheral area was changed. To gain comparable results only one video was used. In a follow-up study the identified good configurations could be used to test different videos, and see whether this is also applicable for other scenes and content.

Still, the proposed system has room for improvement. Saliency models and models of ROI can be incorporated. This can lead to a system that is combining ROIAS and and eye gaze-based video streaming. Thus, the lack between moving gaze and the newly calculated video can be reduced, given participants look at the pre-calculated ROI if this is changing. In order, to improve the perceived quality of the observer, the bit rate difference between the focused and non-focused area can be implemented smoother. That could be done e.g. using several concentric circles with linear or logarithmic increments of lower bit rate, as proposed e.g. in [8].

In order to accommodate for the reported delay between the two differently encoded parts, the used algorithms need to be improved. Furthermore, using faster hardware components could reduce this issue already.

6. Acknowledgements

The work described in this paper has been part of the Bachelor Thesis of Henrik Hesslau [12] and Phong Le Trung [14] who we would like to thank for the technical implementation and running of the experiments.

7. References

[1] 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," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 1–6.

[2] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet." *IEEE Multimedia*, no. 18, pp. 62–67, 2011.

[3] S. Egger, B. Gardlo, M. Seufert, and R. Schatz, "The impact of adaptation strategies on perceived quality of http adaptive streaming," in *Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming*. ACM, 2014, pp. 31–36.

[4] S. Möller, *Quality engineering: Qualität kommunikationstechnischer Systeme*. Springer-Verlag, 2010.

[5] G. Muntean, G. Ghinea, and T. N. Sheehan, "Region of interest-based adaptive multimedia streaming scheme," *Broadcasting, IEEE Transactions on*, vol. 54, no. 2, pp. 296–303, 2008.

[6] B. Watson, N. Walker, L. F. Hodges, and A. Worden, "Managing level of detail through peripheral degradation: Effects on search performance with a head-mounted display," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 4, no. 4, pp. 323–346, 1997.

[7] B. Ciubotaru, G. M., and G. Ghinea, "Objective assessment of region of interest-aware adaptive multimedia streaming quality," *Broadcasting, IEEE Transactions on*, vol. 55, no. 2, pp. 202–212, 2009.

[8] B. Ciubotaru, G. Ghinea, and G.-M. Muntean, "Subjective Assessment of Region of Interest-Aware Adaptive Multimedia Streaming Quality," *Broadcasting, IEEE Transactions on*, vol. 60, no. 1, pp. 50–60, 2014.

[9] J.-S. Lee and T. Ebrahimi, "Perceptual video compression: A survey," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 684–697, 2012.

[10] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[11] R. L. DeValois, B. K. K. DeValois *et al.*, *Spatial vision*. Oxford University Press, 1988, no. 14.

[12] H. Hesslau, "Looking at the perceived Quality of real-time Video Compression with live Eye-Tracking Data," 2014, Technische Universität Berlin, Germany. Bachelor thesis. unpublished.

[13] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," Geneva, International Telecommunication Union, 1998.

[14] P. Le Trung, "Einfluss von unterschiedlich stark komprimierten Videos am Rande des fovealen Sehbereichs auf die Qualitätswahrnehmung," 2015, Technische Universität Berlin, Germany. Bachelor thesis. unpublished.