https://doi.org/10.2352/ISSN.2470-1173.2018.12.IQSP-233

Using The Immersive Methodology to Assess The Quality of Videos Transmitted in UDP and TCP-Based Scenarios

Helard Becerra Martinez[*]*, and Mylène C.Q. Farias*[*+]*;*
[*] *Department of Computer Science,* [+] *Department of Electrical Engineering; University of Brasília, Brasília, Brazil*

Abstract

In this work, we present the results of a psycho-physical experiment in which a group of volunteers rated the quality of a set of audio-visual sequences. The sequences had up to three types of distortions: video coding, packet-loss, and frame freezing distortions. The original content used for the experiment consisted of a set of high definition audio-visual sequences. Impairments were only inserted into the video component of the sequences, while the audio component remained unimpaired. The objective of this particular experiment was to analyze different types of source degradations and compare the transmission scenarios where they occur. Given the nature of these degradations, the analysis is focused on the visual component of the sequence. The experiment was conducted using the basic directions of the immersive experimental methodology.

Introduction

Recent advances on smarthphones technology have transformed services like video conference (Skype, Google Hangout, Facebook Video, FaceTime) and on-demand streaming media (Netflix, iTunes, Hulu, Amazon) into essential tools for the common user. In fact, video applications accounts for the majority of today's internet traffic. Nevertheless, it is understood that the success of these kind of services relies on its trustworthiness and the Quality of Experience (QoE) of the provided service [1].

QoE measures take into account (in addition to Quality of Service features) characteristics of the Human Visual System (HVS) and the Human Auditory System (HAS). Over the last years, several objective quality metrics have been proposed for digital TV [2], lower-resolution video [3], speech [4], and audio signals in general [5]. The videos are delivered to the user using either Transport Control Protocol (TCP) or User Datagram Protocol (UDP) networking protocols. The performance of objective quality metrics is gauged by measuring their correlation with the human perception of quality. The perceived quality is assessed by carrying out subjective experiments, where a group of human participants is asked to rate the quality of a series of signal stimuli (audio, video, or audio-visual sequences) using a particular scale. Recommendations for conducting subjective experiments have been published by telecommunication agencies (International Telecommunications Union ITU, European Broadcasting Union EBU)[2, 3, 4]. Although these experimental recommendations are widely accepted and used, they have limitations in representing the user experience. Several researchers have either modified these methods or proposed new methods to overcome these limitations. Among these methods is the Immersive Methodology proposed by Pinson et al. [6], which puts the human participant in a more natural scenario with the goal of obtaining more realistic subjective measures.

The immersive methodology is specially tailored for multimedia applications that require longer sequences for a better analysis, a type of application in which traditional methodologies have limitations. For example, Garcia *et al.* showed the importance of using an immersive methodology to measure the quality of long videos in adaptive streaming applications [7]. Moreover, Robitza *et al.* used the immersive methodology to study the impact of quality variations and stalling events[8]. Although this experiment used 66 1-minute long source sequences, leading to experimental sessions of over an hour, results showed that the participants's alertness was not affected. Finally, Staelens *et al.* obtained good results using the immersive methodology to perform an experiment that included camera angle changes [9].

In this work, we present the results of a subjective experiment performed using the immersive methodology. In this experiment, a group of human observers rated the audio-visual quality of a set of video sequences containing up to three types of distortions: video coding, packet-loss, and frame freezing impairments. Given the nature of these type of distortions, two groups of test conditions were considered. The first group combined video coding and packet-loss distortions, while the second group considered video coding and frame freezing distortions. The experiment was conducted with the intention of: 1) testing and validating the robustness of subjective data gathered using the immersive methodology, 2) studying and exploring the impact that different types of impairments on perceived audio-visual quality, and 3) producing and publishing a database of audio-visual sequences with a variety of video degradations.

Immersive Methodology

The immersive methodology, proposed by Pinson et al. [6], has the goal of capturing a better estimate of the perceived quality by putting the subject in a more natural scenario. In order to reproduce such scenario, the methodology uses longer stimuli that allows capturing the user's attention. In order to transmit an entire idea and capture the subject's attention, while still maintaining an acceptable test session duration, it is recommended to use 30 to 60 seconds sequences.

The immersive methodology also recommends using audio-visual stimuli, even when video-only or audio-only impairments are being evaluated. The main objective is to maintain a certain level of naturalness in the experimental session. Using audio-visual stimuli has certain consequences. For instance, in an immersive test, subjects are asked to rate the overall audio-visual quality. Beerends and Caluwe [10] showed that participants had trouble separating audio quality from video quality when audio-visual stimuli was presented. The impact of audio quality on

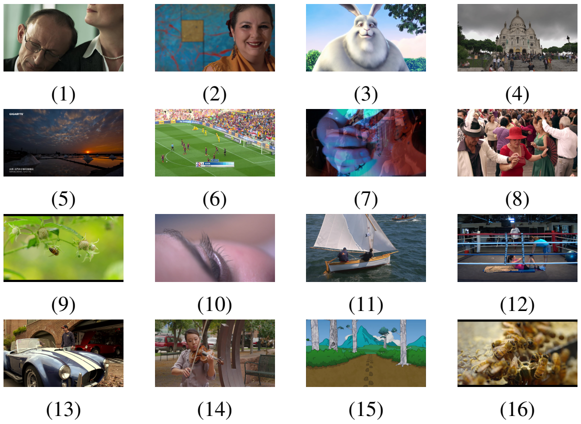


Figure 1: Sixteen sample frames out of the 60 original videos used in the subjective experiment.

video quality (and backwards) can be controlled by evaluating impairments for one component while keeping the quality of the other component constant.

Immersive experiments aim to reduce participants' fatigue during experimental sessions. Traditionally, subjective experiments use a large set of stimuli (audio-visual sequences) processed at a number of Hypothetical Reference Circuit (HRC) with very low content diversity. Presenting these sequences to subjects to assess the quality leads, indefectibly, to boredom and stimuli memorization. In the immersive methodology, each source stimulus (content) is presented only once to each subject. This strategy prevents fatigue and assures that results are not influenced by stimulus memorization.

The basic setup on an immersive experiment is given by a number of source stimuli (w), a set of HRC (y), and a number of subjects (n). The combination of every source stimuli and HRC results in a total of $w \cdot y$ stimuli. For each HRC, Each subject rates w/y of these stimuli. When all subject scores are pooled, approximately n/y subjects rate each individual stimuli.

One last consideration refers to the type of questions made during the experimental task, i.e. after each test sequence is presented. Besides the traditional question regarding the overall perceived quality, the participant is asked to give its opinion about the content. Although content questions are not mandatory in an immersive test, they have the goal of determining whether or not the stimuli is acceptable for a particular application. Also, this type of questions allows investigating the influence of the stimuli content on the perceived quality.

Experimental Setup

Sixty (60) high-definition video sequences (with accompanying audio) were used for this experiment. These sequences were 22 to 68 seconds long, with an average length of 37 seconds. All video sequences have a spatial resolution of 1280×720 (720p), a temporal resolution of 30 frames per second (fps), and a color space format of 4:2:0. Besides the 60 videos used in the experiment, two additional videos were included for trial and training sessions. Figure 1 shows a set of 16 sample video frames taken from the entire set of videos.

Table 1: Parameter Combinations for Freezing Distortions.

Degradation Level		Events	Pos1	Pos2	Pos3	Len1	Len2	Len3
Low	S1	1	2			2		
Medium	S2	2	1	3		1	3	
	S3	2	2	3		2	2	
High	S4	3	1	2	3	2	2	3
	S5	3	1	2	3	3	3	2

Test sequences had impairments caused by: 1) video compression with different codecs at different bitrate levels (e.g. blockiness, blurriness, and ringing.), 2) transmission errors (packet-losses), and 3) transmission caused delays (frame freezing). To generate compression artifacts, we selected two coding standards to compress the source stimuli: H.264/MPEG-4 Advance Video Coding (AVC) and H.265 High Efficiency Video Coding (HEVC) [11, 12]. Four bitrate levels were chosen for each coding standard. To select these values, we visually examined videos compressed at several bitrate levels and choose 4 clear quality levels, taking into account bitrates used in previous work [13, 7].

To generate packet-loss artifacts, the videos were first encoded using H.264 and H.265 codecs and, then, Network Abstraction Layer (NAL) packets were discarted from the video bitstream, similarly to what was done in previous works [14]. Five packet-loss ratios were considered for this experiment: 1%, 3%, 5%, 8%, and 10%.

To create frame freezing distortions, we varied three parameters: number, position, and length of the freezing events. Each video sequence were likely to have one, two, or three freezing events. We chose three freezing positions, which are labeled as '1', '2' and '3'. The positions were determined dividing the total length of the video sequences by three and multiplying the result by: 0, 1, and 2. Therefore, a freezing position '1' represents an initial loading distortions, which is experienced before the video starts playing, while a freezing positions '2' and '3' represent freezing in the second and third slots. Finally, the length of the freezing events were fixed at '1', '2', and '3' seconds.

All three parameters (number, position, and length of the freezing event) were combined to vary the level of the overall degradation and, consequently, of the level of annoyance perceived by the user. The levels were set as 'S1', 'S2', 'S3', 'S4' and 'S5', going from the least annoying combination (S1) to the most annoying combination (S5). Table 1 depicts the parameters corresponding to the combinations.

Since frame freezing and packet-loss distortions do not occur simultaneously in a real transmission scenario [7], we split the HRCs in two groups. The first group combines artifacts produced by compression with packet-loss distortions (HRC1 to HRC5). The second group combines artifacts produced by compression with frame freezing distortions (HRC6 to HRC10). Additionally, two video sequences compressed at extremely high bitrate levels, with no packet-loss video distortions or frame freezing effects, were used as anchors to help participants establish the entire range of quality used for the experiment.

For the first HRC group, five combinations of bitrate levels and codecs were chosen, representing five levels of quality. For each of these combinations, packet-loss distortions were in-

Table 2: First group of HRCs.

HRC	Codec	Bitrate (kb/s)	PLR
HRC01	H.264	500	10%
HRC02	H.265	400	8%
HRC03	H.264	2,000	5%
HRC04	H.265	1,000	3%
HRC05	H.265	8,000	1%

Table 3: Second group of HRCs.

HRC	Codec	Bitrate (kb/s)	Freezing
HRC06	H.265	200	S5
HRC07	H.264	800	S4
HRC08	H.265	1,000	S3
HRC09	H.264	2,000	S2
HRC10	H.264	16,000	S1

troduced at 5 different ratios (1%, 3%, 5%, 8%, and 10%). Table 2 shows the resulting HRCs. These 5 HRCs were replicated for all 60 source stimuli, resulting in three hundred (300) test stimuli.

For the second HRC group, another five combinations of bitrate levels and codecs were used. It is worth mentioning that no combination used for the first group was used for the second group. Each of these five encoding combinations was paired with one of the five levels of frame freezing (S1, S2, S3, S4, and S5). Table 3 displays the resulting combinations. These 5 HRCs were replicated for all 60 source stimuli, resulting in three hundred (300) test stimuli.

To avoid a saturation of the rating scales (quality and content), two anchors were used. These anchors were encoded using H.264 and H.265 codecs at high bitrate levels. This procedure was performed for all 60 source stimuli, resulting in one hundred and twenty (120) anchors. Pooling all test stimuli, seven hundred and twenty (720) test videos were generated for this experiment. Following the Immersive Methodology, each participant was presented with 60 test stimuli. It is worth pointing out (again) that each participant watched only one test sequence generated from a specific original.

Experimental Methodology

The experiment was carried out in the recording studio of the Núcleo Multimedia e Internet (NMI). NMI is part of the Department of Engineering (ENE) at the University of Brasilia (UnB). The experiment was run with one subject at a time, using a desktop computer, a LCD monitor, and a set of earphones. The subjects were seated straight ahead of the monitor, centered at or slightly below eye height for most subjects. The distance between the subjects eyes and the video monitor was set at three screen heights, which is a conservative viewing distance according to the ITU-T Recommendation BT.500.1 [15].

The experiment was performed with 60 participants ($n = 60$). As mentioned before, recommendations presented in the Immersive Methodology [6] were used for this experiment. The test was divided into three main sessions: Overview, Training Session, and Main sessions. For the Overview session, participants were presented with a set of original source videos and a set of corresponding degraded versions (HRC combinations). The objective of this session was to familiarize the participant with the quality range of the test sequences in the experiment.

In the Training session, subjects performed the same tasks performed in the main session. The goal of the training session was to expose subjects to sequences with impairments and give them a chance to try out the data entry procedure. In the main session, the actual experimental task was performed. A break was introduced in the middle of the main session to allow the subjects to rest.

After observers watched with each test sequence, they were asked two questions. The first question was about the overall audio-visual quality of the test sequence, while the second question was about the video content. To answer these questions, participants were presented with a five point Absolute Category Rating (ACR) scale ranging from '1' to '5'. For the audio-visual quality question, the scale was labeled (in Portuguese) as "Excellent", "Good", "Fair", "Poor", and "Bad". For the content question, the scale was labeled as "Intriguing", "Interesting", "Neutral", "Uninteresting", and "Boring". Scale labels were taken from the immersive speech quality test conducted in [6]

Traditionally, the mean opinion score (MOS) for each test video is obtained by taking the average of the subjective scores given to each test video, by all observers. In our experiment, two different subjective scores where gathered: *quality* and *content* scores. The Mean Quality Score (MQS) per-HRC is obtained by averaging the quality scores given by all subjects:

$$\mathrm{MQS}_{\mathrm{HRC(j)}} = \frac{1}{n} \cdot \sum_{i=0}^{n} QS_j(i), \tag{1}$$

where n is the total number of subjects and $QS_j(i)$ is the quality score given by the i-th subject to the j-th HRC test sequence, with $j = \{1, 2, \ldots, 12\}$. In other words, $\mathrm{MQS}_{\mathrm{HRC(j)}}$ gives the average quality score for the j-th HRC, measured over all subjects and originals. Similarly, the Mean Content Score (MCS) per-HRC is obtained by taking the average of the content scores given by all subjects:

$$\mathrm{MCS}_{\mathrm{HRC(j)}} = \frac{1}{n} \cdot \sum_{i=0}^{n} CS_j(i), \tag{2}$$

where $CS_j(i)$ is the content score given by the i-th subject to the j-th HRC test sequence, with $j = \{1, 2, \ldots, 12\}$.

Results: Internal Consistency

As a starting point, the confidence levels are calculated for each of the average subjective scores (MQS and MCS). A high score variability may indicate a low confidence level and, therefore, a low reliability of the results. To evaluate the reliability of the results, we calculate the Cronbach's α coefficient [16] for the MQS and MCS values (see Table 4). The α coefficient values range from 0 to 1, with higher values implying a greater internal consistency (reliability). More specifically, α coefficients with values from 0.00 to 0.69 correspond to a 'poor' internal consistency, values from 0.70 to 0.79 correspond to a 'fair' internal consistency, values from 0.80 to 0.89 correspond to a 'good' internal consistency, and values from 0.90 to 1 correspond to an 'excellent' internal consistency.

Table 4: Cronbach's α for all MQS_{HRC} and MCS_{HRC}.

Score	Analysis	Cronbach's α
MQS_{HRC}	per-HRC	0.924
MCS_{HRC}	per-HRC	0.858

As shown in Table 4, for the per-HRC analysis, the α coefficient for MQS_{HRC} is 0.924, while the α coefficient for MCS_{HRC} is 0.858. This suggests that, for the different HRCs, subjects agreed more on the quality score than on the content score. Also, these results show that the MQS and MCS values gathered are highly reliable, making the immersive experimental methodology very reliable.

Experimental Results

As described earlier, HRCs were divided into two groups. The first group includes HRCs from 1 to 5 (including the anchor 1), which corresponds to test sequences containing coding impairments and distortions due to packet-loss (see Table 2). The second group includes HRCs from 6 to 10 (including the anchor 2), which corresponds to test sequences containing coding impairments and frame freezing distortions (see Table 3).

Coding/Packet-Loss Scenario

Figure 2 depicts MQS_{HRC} values, including a 95% confidence interval, for the first group of HRCs (coding/packet-loss scenario). Each HRC corresponds to a combination of bitrate level (BR) and packet-loss ratio (PLR) values, as detailed in Table 2. Notice that MQS_{HRC} increases for most BR and PLR combinations. Yet, the difference between HRC01 and HRC02 MQS_{HRC} values is very small (no statistical significance). A similar behavior is observed between HRC3 and HRC4. MQS_{HRC} values range from 1.95 to 4.30, with no evidence of a scale saturation. This suggests that participants were able to distinguish between the different levels of impairments used in this scenario.

Figure 3 depicts the MQS_{HRC} as a function of the packet-loss ratio values (PLR). In this figure, the H.264 MQS_{HRC} values are shown in blue, while the H.265 MQS_{HRC} values are shown in green. It can be observed that MQS_{HRC} values drop as the PLR is increased and the bitrate is decreased. However, very similar MQS_{HRC} values (no statistical difference) are observed for two different cases: HRC04 (PLR = 3% , BR = 1000kb/s, and Codec = H.265) and HRC03 (PLR = 5% , BR = 2000kb/s, and Codec = H.264). Previous studies [17, 13] show that a video encoded with H.264 at a certain bitrate has approximately the same quality as a video encoded with H.265, but with half of the bitrate. In Figure 3, this can be observed for HRC04 and HRC03 (1000kb/s, H.265 and 2000kb/s, H.264).

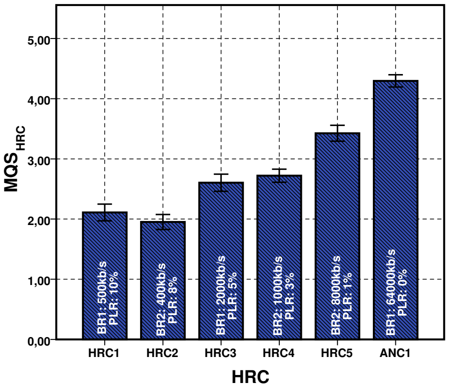


Figure 2: MQS_{HRC} for the coding/packet-loss scenario. See HRC specifications in Table 2.

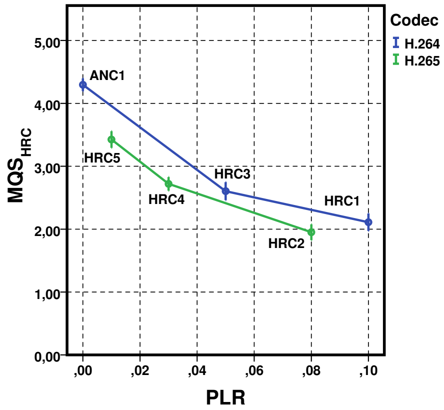


Figure 3: MQS_{HRC} versus packet-loss ratio (PLR). See HRC specifications in Table 2.

It is worth pointing out that HRC04 and HRC03 have different packet-loss rate values (3% and 5%). This might indicate that the coding algorithms responded differently to packet-loss distortions. From the literature [18, 19], it has been shown that H.265 is more sensitive to packet-losses than H.264. This might explain why the MQS_{HRC} corresponding to a higher PLR (5% for HRC03) is not statistically different of the MQS_{HRC} corresponding to a lower PLR (3% for HRC04). On the other hand, the MQS_{HRC} difference between HRC02 (PLR = 8% , BR = 400kb/s, and Codec = H.265) and HRC01 (PLR = 10% , BR = 500kb/s, and Codec = H.264) is statistically significant. In this case, since H.265 is more sensitive to packet-loss, this artifact had a greater effect on quality than the compression bitrate.

For this first scenario, it has been observed that the video bitrate, the coding algorithm, and the PLR all have an important impact on the perceived audio-visual quality (MQS_{HRC}). However, for certain rates of packet-loss, the coding algorithm is proven to be more determinant.

Coding/Freezing Scenario

Figure 4 presents the MQS_{HRC} values, including a 95% confidence interval, for the coding-freezing scenario. Each HRC corresponds to a combination of a bitrate level (BR) and a frame freezing level, which is given by the number of pause events (N), the position of the pause event (P), and the length of the pause events (L). Table 1 presents all frame freezing distortion levels used in this work. Notice that MQS_{HRC} increases with the bitrate (BR) level and decreases with the number of pause events (N). This behavior pattern is observed for all HRCs. MQS_{HRC} val-

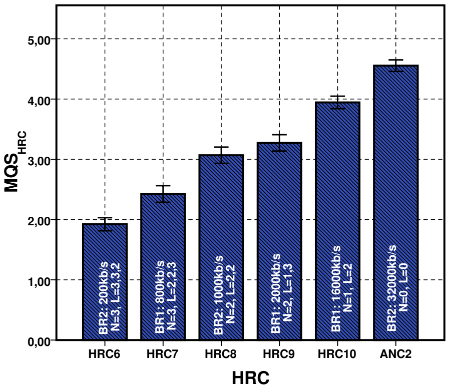


Figure 4: $\mathrm{MQS}_{\mathrm{HRC}}$ for the coding-freezing scenario. See HRC specifications in Table 3.

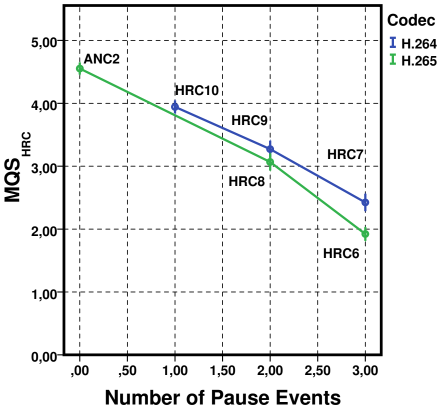
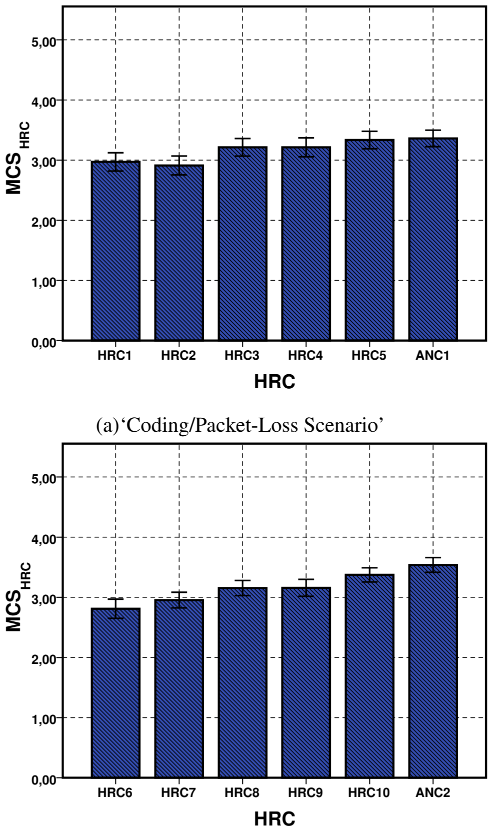


Figure 5: $\mathrm{MQS}_{\mathrm{HRC}}$ according the Number of pause events. See HRC specifications in Table 3.



(a)‘Coding/Packet-Loss Scenario’

(b) ‘Coding-Freezing Scenario’

Figure 6: $\mathrm{MCS}_{\mathrm{HRC}}$ for both scenarios.

ues range from 1.92 to 4.55, with no sign of a scale saturation. This suggests that participants were able to distinguish between the different levels of impairments used for this scenario.

Figure 5 presents the $\mathrm{MQS}_{\mathrm{HRC}}$ as a function of the number of pause events (N). The figure presents the different HRCs, for both H.264 and H.265 codecs. For the particular case of HRC08 and HRC09 (same number of pause events), it can be inferred that the $\mathrm{MQS}_{\mathrm{HRC}}$ difference was determined by the position (P) and length (L) of the pause events, since a certain equivalence is expected in terms of bitrate [17, 13].

For HRC09, the pause events were located at positions ‘1’ and ‘3’, and their durations were 1 and 3 seconds respectively. For HRC08, the pause events were located at positions ‘2’ and ‘3’, and their durations were 2 seconds. By comparing these values, we can see that a short pause at the beginning of the playout (initial loading) is less annoying than a pause during the playout. This result is in accordance with previous studies [20]. For the case of HRC06 and HRC07 (same number of pause events), the higher $\mathrm{MQS}_{\mathrm{HRC}}$ difference might be attributed to their bitrate levels (200kb/s and 800kb/s) and the positions and durations of the pause events. For HRC06, the pause events were located at positions ‘1’ , ‘2’ and ‘3’, and their durations were 3, 3, and 2 seconds, respectively. For HRC07, the pause events were located at positions ‘1’ , ‘2’ and ‘3’, and their durations were 2, 2, and 3 seconds, respectively.

Based on these results, we conclude that there is an additive impact of frame freezing and video bitrate on $\mathrm{MQS}_{\mathrm{HRC}}$. The impact of frame freezing can be determined by the number, position, and duration of pause events.

Content Score Results

Figures 6 (a) and 6 (b) present the $\mathrm{MCS}_{\mathrm{HRC}}$ values for each HRC, corresponding to both coding/packet-loss and coding/freezing scenarios. It can be observed that, for both scenarios, the range of $\mathrm{MCS}_{\mathrm{HRC}}$ is much smaller than the range of $\mathrm{MQS}_{\mathrm{HRC}}$, fluctuating around ‘3’ (“Neutral” Content). Judging by the small difference presented among all HRCs and ANCs, it might be inferred that subjects didnt perceived mayor differences between the degraded sequences and the originals in terms of content.

For a better comparison between quality and content scores, Figure 7 shows a barplot of $\mathrm{MQS}_{\mathrm{HRC}}$ superimposed with a graph of $\mathrm{MCS}_{\mathrm{HRC}}$, for all HRCs. Although the $\mathrm{MCS}_{\mathrm{HRC}}$ range is smaller, it is possible to observe that it varies with the perceived quality. This behaviour suggests that the participant’s opinion about the content is somewhat related to its perceived quality. Or, the perceived quality seems too be influenced by the video con-

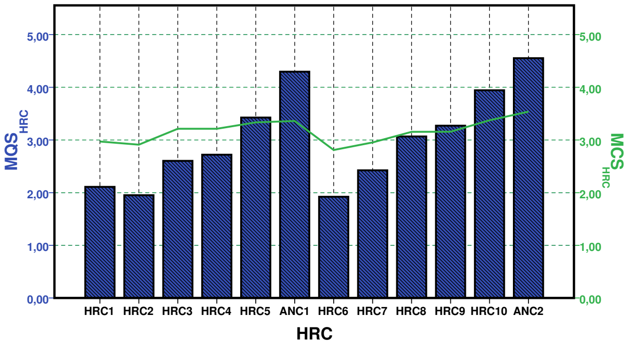


Figure 7: Evolution of both MQS_{HRC} and MCS_{HRC} scores along all HRCs.

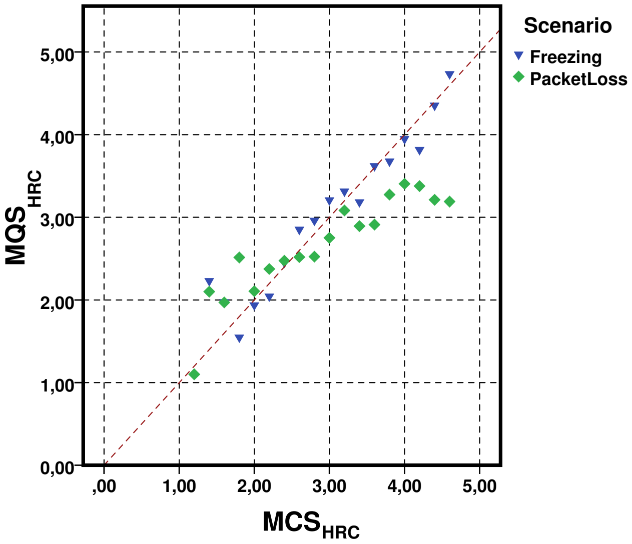


Figure 8: Scatter plot comparing subjective results of MQS_{HRC} and MCS_{HRC} for the two scenarios.

tent [21, 22]. In spite of the mutual impact of content and quality, further experiments are needed to better understand this behavior.

Figure 8 depicts a scatter plot of the MQS_{HRC} and MCS_{HRC} values for the coding/packet-loss and coding/freezing scenarios. Notice that participants gave higher MQS_{HRC} and MCS_{HRC} values to test sequences in the coding-freezing scenario. So, participants were more tolerant to pauses during the video playout than to severe visual distortions (blocking, slicing, blockloss) caused by packet losses. Such tolerance is also reflected on the MCS_{HRC} values, suggesting that the participant's opinion of the content is affected by the distortions.

Conclusions

This paper presented a subjective video quality experiment conducted using the immersive experimental methodology. The experiment contained audio-visual sequences impaired with three types of degradations: video coding, packet-loss, and frame freezing impairments. Given the nature of these degradations, two test scenarios were considered: a coding/packet-loss scenario and a coding/freezing scenario. Five HRCs (and one anchor) were used in each scenario, resulting in 12 different levels of distortions. These levels of distortion were replicated for the 60 audio-visual original sequences, producing 720 test sequences. A total of 60 participants performed the experiment. Each participant watched 60 video sequences, of different content, and was asked to score the quality and the content of each of these sequences.

To verify the consistency of the subjective data and, therefore, the reliability of the immersive methodology, the Cronbach's alpha coefficient was calculated for the quality and content scores. Results showed a good, near excellent, level of consistency. These results validate the use of the immersive methodology. Regarding the content scores, results suggest that participant's opinion of the content was affected by its perceived quality.

The per-HRC analysis of the coding/packet-loss scenario showed that the perceived quality is affected by packet-loss and bitrate. Results also shows the sensitivity of the H.265 codec to packet-loss impairments. For the coding/freezing scenario, the perceived quality was affected by the number, duration, and position of pause events. In general, results showed a smaller tolerance to frame freezing distortions, when compared to packet-loss distortions. In other words, visual degradations were more annoying than freezing degradations.

Acknowledgments

This work was supported in part by the Conselho Nacional de Desenvolvimento Cientfico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Fundação de Apoio a Pesquisa do Distrito Federal (FAP-DF), and the University of Brasília (UnB).

References

[1] J. Korhonen, "Audiovisual quality assessment in communications applications: current status, trends and challenges," *Signal Processing*, pp. 6–9, 2010.

[2] I. ITU, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference."

[3] I. T. S. Sector, "Objective perceptual multimedia video quality measurement in the presence of a full reference," *ITU-T Recommendation J*, vol. 247, 2008.

[4] I. Rec, "G. 107-the e model, a computational model for use in transmission planning," *International Telecommunication Union*, vol. 8, pp. 20–21, 2003.

[5] ITU-R, "Recommendation bs.1387 : Method for objective measurements of perceived audio quality," Tech. Rep., 1998.

[6] M. Pinson, M. Sullivan, and A. Catellier, "A new method for immersive audiovisual subjective testing," in *Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 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," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 129–134.

[8] Robitza, Werner and Garcia, Marie Neige and Raake, Alexander, "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pp. 1–6, 2015.

[9] Staelens, Nicolas and Coppens, Paulien and Van Kets, Niels and Van Wallendaef, Glenn and Van den Broeck, Wendy and De Cock, Jan and De Turek, Filip, "On the impact of video stalling and video quality in the case of camera switching during adaptive streaming of sports content, " in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, pp. 1–6, 2015.

[10] F. E. Beerends, John G.; De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc*, vol. 47, no. 5, pp. 355–362, 1999.

[11] ITU-T, "H.264 : Advanced video coding for generic audiovisual services," Tech. Rep., 2003.

[12] ——, "H.265 : High efficiency video coding," Tech. Rep., 2013.

[13] M. Horowitz, F. Kossentini, N. Mahdi, S. Xu, H. Guermazi, H. Tmar, B. Li, G. J. Sullivan, and J. Xu, "Informal subjective quality comparison of video compression performance of the hevc and h. 264/mpeg-4 avc standards for low-delay applications," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2012, pp. 84 990W–84 990W.

[14] J. Redi, I. Heynderickx, B. Macchiavello, and M. Farias, "On the impact of packet-loss impairments on visual attention mechanisms," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1107–1110.

[15] ITU-R, "Recommendation bt.500-8: Methodology for subjective assessment of the quality of television pictures," Tech. Rep., 1998.

[16] J. M. Bland and D. G. Altman, "Statistics notes: Cronbach's alpha," *Bmj*, vol. 314, no. 7080, p. 572, 1997.

[17] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standardsincluding high efficiency video coding (hevc)," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1669–1684, 2012.

[18] B. Oztas, M. T. Pourazad, P. Nasiopoulos, and V. Leung, "A study on the hevc performance over lossy networks," in *Electronics, Circuits and Systems (ICECS), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 785–788.

[19] P. Pinol, A. Torres, O. Lopez, M. Martinez, and M. P. Malumbres, "Evaluating hevc video delivery in vanet scenarios," in *Wireless Days (WD), 2013 IFIP*. IEEE, 2013, pp. 1–6.

[20] 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.

[21] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2d and 3d subjective video quality tests," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–12, 2013.

[22] P. Kortum and M. Sullivan, "The effect of content desirability on subjective video quality ratings," *Human factors: the journal of the human factors and ergonomics society*, vol. 52, no. 1, pp. 105–118, 2010.