DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

Subjective Analysis and Objective Characterization of Adaptive Bitrate Videos

Jacob Søgaard[a], Samira Tavakoli[b], Kjell Brunnström[cd], Narciso García[b]
[a]Technical University of Denmark, Lyngby, Denmark
[b]Universidad Politécnica de Madrid, Madrid, Spain
[c]Acreo Swedish ICT AB, Kista, Sweden
[d]Mid Sweden University, Sundsvall, Sweden

Abstract

The HTTP Adaptive Streaming (HAS) technology allows video service providers to improve the network utilization and thereby increasing the end-users' Quality of Experience (QoE). This has made HAS a widely used approach for audiovisual delivery. There are several previous studies aiming to identify the factors influencing on subjective QoE of adaptation events. However, adapting the video quality typically lasts in a time scale much longer than what current standardized subjective testing methods are designed for, thus making the full matrix design of the experiment on an event level hard to achieve. In this study, we investigated the overall subjective QoE of 6 minutes long video sequences containing different sequential adaptation events. This was compared to a dataset from our previous work performed to evaluate the individual adaptation events. We could then derive a relationship between the overall mean opinion score (MOS) and the MOS from shorter sequences. The aforementioned empirical dataset has proven to be very challenging in terms of video quality assessment test design, thus deriving a conclusive outcome about the influence of different parameters have been difficult. The second contribution of this study is considering how objective characterizations of adapted videos can improve the understanding of the subjective ratings.

Introduction

Video delivery accounts for the major share of nowadays Internet traffic. A large portion of this traffic is delivered through HTTP server-based streaming services such as YouTube, using TCP as underlying transport protocol. In contrast to more traditional video delivery methods over UDP, TCP's packet retransmission property prevents audio/video distortions. However, the delivery channel throughput may vary strongly. When the available bandwidth falls below the video bitrate, the client buffer depletes and the playback is interrupted, resulting in video stalling. To avoid the negative impact of buffering/stalling events on users' Quality of Experience (QoE), streamed videos have to be either encoded at very low bitrates or adapted dynamically to the available bandwidth.

A popular method of video delivery today is online streaming of videos using HTTP Adaptive Streaming (HAS). In this setting, different quality representations of the media is available at the server and partitioned into segments of typically 1 to 10 sec. Depending on the current network throughput, the video playback client chooses the next segment from the available quality representations of the media aiming to reduce the risk of buffer depletion, and therefore stalling events. Nevertheless, switching between different representations will result in a video playback with time-varying quality, which might in itself give rise to a new type of visual degradation.

Bitrate adaptation can be done through different strategies. However, to provide an optimal viewing experience for the end-user, it is crucial to understand the QoE impact of quality variation due to adaptation. The ultimate goal for service providers is to optimize the QoE of an entire viewing session that could be anything from a few minutes up to maybe 1-2 hours. Despite this, the research on the relationship between the overall QoE of (long) HAS video sequences and the perceptual quality of individual adaptation events has not received much attention yet and is remaining an open research question.

On the other hand, evaluation of such a long video sequence is challenging as the existing testing methods are designed for the short test stimuli up to one minute and not being appropriate to evaluate the video quality varying over a longer time. This makes the full matrix design of the experiment on an event level hard to achieve.

As an alternative, one can turn to objective Video Quality Assessment (VQA) methods to get estimations of the video quality of HAS videos. In order to understand the overall subjective results and eventually build a good objective quality estimator, a good understanding of the factors influencing the subjective quality is required. The idea here is that the subjective data can be compared to data objectively characterize the video sequences, at least with objective metric building blocks, e.g. spatial and temporal information, contrast sensitivity and masking, blockiness, blur, etc. This could give valuable insight into how an effective metric should be constructed.

Objective VQA methods can be divided into the following three main categories: Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) quality assessment. In the FR scenario, the original video is available for measuring along with the distorted version. In the RR scenario, only a limited set of information about the original video is available. Typically, relevant parameters are computed and transmitted of the original and or the distorted video. Finally, in the NR scenario no information about the original video is available.

Objective NR VQA methods are very useful since no additional data is transmitted along with the video signal. Thus, the algorithms can be carried out solely at the receiving end and without affecting the encoding or the amount of transmitted data. Objective VQA of HAS videos are still very much an open problem and even more so if the method is required to be a NR method.

The contribution of this study is two-fold. First it investigates the relationship between the QoE of long video sequences including sequential adaptation events and the perceptual quality of

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

containing individual adaptation events. This is done through subjective evaluation of 6 minutes long video sequences including different adaptation events and comparing the obtained results to a dataset from our previous study which targeted the QoE of containing individual adaptation events. The second contribution of the paper is to investigate the relationship between the objective characterization of adaption events and the corresponded subjective ratings.

Related Works

Through an extensive literature review presented in surveys [1], [2] the factors influencing the QoE of HAS services have been presented. For instance, many studies have shown that gradual quality variation is preferred over abrupt one. The frequency of quality representation switches negatively impacts the experienced quality. People seem to prefer constant over time-varying quality, unless the average quality is too low—in that case, any switch to a higher quality is better. Also, it is generally agreed upon that any kind of stalling must be avoided.

In spite of presenting the aforementioned findings, the authors of [1] also address many questions which have remained open or not appropriately solved due to (i) a limited number of tests conducted to address a question, (ii) shortcomings of the reported studies (such as missing information in the respective publication), or evident limitations in the considered set of test conditions, (iii) methodological shortcomings in terms of how tests have been conducted, or (iv) contradictory outcomes with respect to identical research questions.

On the other hand, there are some new research questions, especially in regard to testing methodology to evaluate the QoE for HAS long sequences. The only standardized method aimed at long sequences, Single Stimulus Continuous Quality Evaluation (SSCQE) from ITU-R Rec. BT.500 [3], requires the user to focus on both rating and watching at the same time. However, the recency and hysteresis effect of the human behavioral responses while evaluating the time-varying video quality would lead to an unreliable evaluation through this methodology [4].

In fact, research on new methodologies to improve the current subjective testing approaches has been already started. For instance, a new approach for immersive evaluation of audiovisual content was proposed in [5]. This method is based on the use of long test stimuli to encourage the observers' engagement with the content and simulating real situations of using audiovisual applications. In this work, not only longer sequences are recommended for evaluating video quality, but also using test sequences with audio (in spite of traditional standard recommendations). This recommendation makes sense since video-only presentations poorly represent the users' experience of an audiovisual application, as people rarely watch videos without audio.

Another new approach was presented in our previous work [6], in which the evaluation of set of subsequent adaptation scenarios was made using long sequences (around 6 minutes). The idea behind designing this method, which has been named Content-Immersive Evaluation of Transmission Impairments (CIETI), was to simulate realistic viewing conditions by using longer sequences so the observers become more engaged to the content as they would be in real life, rather than focusing on detecting impairments, which can happen using traditional methodologies with the short test videos. Nevertheless, the voting period in this approach is intrusive and may prevent the user from fully immersing into the content. Thus, understanding the relationship between the QoE of individual adaptation event evaluated during the voting time and the whole video sequence including subsequent adaptation events would be challenging.

There are only few contributions in the field of objective Video Quality Assessment (VQA) research that specifically targets QoE in HAS. As shown in [1] distortion-generic state-of-the-art Image Quality Assessment (IQA) and VQA methods does not achieve very good performance when applied directly to HAS test sequences. The authors of [1] also outline an objective VQA method based on time and frequency of buffering events, the video quality level, frame rate, and the Mean Opinion Score (MOS) of the video streams at the encoder side. In [8] a No-Reference (NR) QoE estimation method for H.264/AVC HAS videos based on quantification of buffering events, the quantization, and mapping by a neural network is presented.

The performance of different temporal pooling methods with the IQA methods Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity index (SSIM) on HAS videos is presented in [9]. Similarly, initial work on a Full-Reference (FR) objective VQA method based on pooling of IQA scores, but also considering different bitrate information is presented in [10].

Despite the promising results in the previous work, good performance on a more realistic HAS dataset can be hard to achieve. Motivated by this, work on objective characterization of HAS videos is presented in this paper, which can be used in the development of more robust objective VQA methods for HAS.

Study Description

Taking the open questions from previous studies [1], [2] into account, following factors about adaptation behaviors were considered for further investigation:

- What is the perceptual effect of technical switching parameters, specifically, abrupt vs. smooth switching and chunk size when decreasing and increasing the video quality?
- What is the influence of content type on perceptual quality of switching strategies?
- Is it better to switch the quality level or try holding a certain (even low) quality level to minimize the impairment caused by the switching itself?
- What is the proper methodology to subjectively evaluate the adaptive streaming strategies? What is the influence of audio presence on evaluation of video-related impairments?

To study these research questions, we previously performed three subjective experiments considering various adaptation scenarios and using different testing methodologies to evaluate the individual adaptation events [11]. In current study, we conducted a new experiment to evaluate the overall QoE of the whole long video sequence. This experiment is presented in the section of Subjective Experiment in the following section.

In the design of our previous study, special care was taken in order to provide a realistic HAS system environment. Nevertheless, due to the high number of sequences and their duration, applying all test scenarios on each video source was not possible which made it difficult to determine whether the difference in subjective scores arises from the different type of distortions or the different video sources. Therefore, in this study,

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

we consider objective characterization of the video content to get a better understanding of the dataset.

Subjective Experiment

Test Materials and Conditions

Among commercial content, seven source video sequences (SRC) of approximately 6 min long were chosen as listed in Table 1. The spatial and temporal activities of the content, which was determined using the metric provided by [18], covered a large portion of SA-TA plane.

Table 1: Characterization of Source Video Sequences (SRC) used in the experiments

Code	Type	Format	Description
P	Movie	1080p 24fps	Action, with some scene in smooth motion, some with group of walking people, some with camera panning
S	Movie	1080p 24fps	Drama, romance, mostly with the smooth motion in the static background, some scene with group of dancing people in bright ambient
D	Movie	1080p 24fps	Action, Si Fi, with the rapid changes in some sequences, cloudy atmosphere
C	Documentary	1080p 50fps	Sport documentary, mostly with handheld shooting camera
F	Sport	1080p 50fps	Soccer, average motion, wide angle camera sequences, uniform camera panning
N	News	1080p 50fps	Spanish news, some scenes with static shooting camera with one/two standing/sitting people; some outdoor scenes, other scenes with camera pan
R	Music	1080p 50fps	Music concert, high movement of the singer with some sudden scene change

The video representations were provided considering the compression domain as switching dimension and quality range used in practice for the living-room platform. For each SRC, four quality levels from 600 kbps to 5 Mbps were produced using Rhozet Carbon Coder with the setting summarized in Table 2. It was assumed that the network bandwidth varies along these levels.

Table 2: Transcoding parameters of adaptive streams' quality levels

Code	Frame rate	Resolution	Target bitrate (kbps)
Q1	24	720p	600
Q2	24	720p	1000
Q3	24	720p	3000
Q4	24	720p	5000
Video: H.264, high profile, closed GoP, disabled scene change detection Ref. frame: 2, B frame: 2, Constant Bitrate Coding (CBR), Adaptive QP			
Audio: AAC, 192 Kbps			

For each of the status when client should request from the server lower bitrate chunk (down-switching) or higher bitrate chunk (up-switching), four Hypothetical Reference Circuits (HRC) were constructed including abrupt and smooth switching each using two different chunk lengths. For the comparison of abrupt versus smooth switching strategies, the two video sequences to be compared against each other shared the same lower, Q_i, and higher quality level, Q_{i+k}, (cf. *code* in Table 2), with i indicating the lower quality level of the respective sequence and k indicating the number of quality level change for reaching to the higher one. In the case of abrupt switching, the quality change occurred in the middle of sequence duration, whereas for the smooth switching after every chunk one quality change took place until reaching to the target level. Since human perception of quality switching can be different with respect to the switching direction, abrupt and smooth switching test sequences were constructed for both up and down-switching directions. For each of these switching behaviors, two chunk size, 2 sec and 10 sec length, were considered to be inline with current HAS solutions. To study the perceptual quality of adaption streams in different content, four HRCs were considered representing the constant quality level. The list of all HRCs is presented in Table 3.

Table 3: List of the test adaptation Strategies (HRCs)

Status	Possible Behavior		Code
Increasing quality	Gradual change	10 s chunk	IGR10
		2 s chunk	IGR2
	Rapid change	10 s chunk	IRP10
		2 s chunk	IRP2
Decreasing quality	Gradual change	10 s chunk	DGR10
		2 s chunk	DGR2
	Rapid change	10 s chunk	DRP10
		2 s chunk	DRP2
Constant quality	Whole sequence at 5 Mbps		N5
	Whole sequence at 3 Mbps		N3
	Whole sequence at 1 Mbps		N1
	Whole sequence at 600 kbpd		N600

To produce the test sequence (TS), each SRC was segmented following the pattern shown in Figure 1. Because of the session time limitation and high number of SRCs and HRCs, the full factorial design was not feasible. To respect the ITU recommended test session length [3], four out of seven SRCs (P, S, F and C content- cf. *code* in Table 1) were selected to be prepared in two different variants. By means of these two variants (called as 'content code'-V1 and 'content code'-V2 in the Results section), relevant switching behaviors (i.e. comparing GRx and RPx) as well as the constant quality HRCs with potential non-perceivable difference (i.e. comparing N3 and N5, as well as N600 and N1) were compared in an identical segment of the aforementioned content. As a result, 11 TSs, i.e. for each HRC, 11 different individual segments (4x2+3), and consequently the total of 132 Processed Video Sequences (PVS) (11x12) were generated for evaluation. Length of the PVSs was variable depending on the HRCs: 40 sec for those considering the quality switching with 10 sec chunk (cf. xGR10 and xRP10), and 14 sec for rest of the HRCs.

Figure 1: Format of test sequence (TS) according to CIETI methodology. PVS and VS stand on 'processed video sequence' and 'voting segment' in order. '0' printed in the corner of the first segment's frames has no degradation indicating the start of the test. In the test session, randomized order of test sequences were presented to the subjects.

Evaluation Approach

Previous study

In our previous study presented in [11], three experiments were conducted in different environments and through different testing approaches to evaluate identical PVSs.

The first experiment was conducted in Acreo Swedish ICT's lab (denoted as `Acreo' experiment). The randomized order of all

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

the PVSs (cf. PVS in Figure 1) were presented to the test subjects following the Absolute Category Rating (ACR) methodology adapted from ITU-T Rec. P.910 [18]. After presentation of each PVS, the test subjects were asked to answer two questions pop-up on the screen: overall quality of the PVS (rating on the _five-graded ACR scale (Bad, Poor, Fair, Good and Excellent which was later mapped to the scores 1, 2, 3, 4 and 5 respectively), and if they perceived any change in the quality (options: Increasing, No change, Decreasing).

The other two experiments were carried out in the Universidad Politécnica de Madrid's (UPM) lab using CIETI method: one by presenting only the video stimulus (denoted as `UPM-NoAudio' experiment), and the other one in the presence of audio in constant quality (denoted as `UPM-Audio' experiment). In the test session, the 11 TS each including 12 sequential PVS-VS pairs (cf. Figure 1) were presented in a randomized order. For the evaluation, the test subjects were asked to answer the same questions as in the Acreo experiment and using identical rating scales but on paper questionnaires instead. As a new task, after evaluating the 12 PVSs of each TS, there was another question in the questionnaire asking about the overall quality of the whole sequence. 40 sec after terminating the evaluation of each TS, the next one was played.

In order to allow for cross-lab comparison, the ambient and all the hardware and software in Acreo were adjusted similar to the UPM complying with the ITU-R Rec. BT.500-13 [3]. A 46" Hyundai S465D display was used with the native resolution of 1920 x1080. The resolution of the TV was set to 1280x720 to present the video in full screen, thus scaling was performed by the TV. The viewing distance was set to four times the display height. The TV's peak white luminance was 177 cd/m^2 and the illumination level of the room was 20 lux.

Current study

The current experiment was conducted in Acreo's laboratory under the same settings as previous experiments but the test stimuli to be evaluated was considered the entire 6 min video sequence including the individual adaptation events (see Figure 1).

30 test subjects participated in the Acreo experiment. There were 20 male and 10 female test subjects with age range of 12 to 63, with various backgrounds. All test subjects had English as second language, their mother tongue varied between Swedish and various languages. However, the majority had Swedish as their native language. The written instructions, voting scale and pre- and post-questionnaires were given in English. The conversation and questions between test leader and test subject were either Swedish or English depending on the preference of the test subject.

On arrival, the test subjects were handed the written instructions and when they had been carefully read. A visual test was then performed for visual acuity (Snellen) and color vision (14-plate Ishihara). Their performances were noted down. Nobody was screened based on this. All participants except one had at least 0.8 on at least one eye and perfect color vision (all plates correctly answered on our Ishihara test). No test subject was screened from the test. The test subjects were then given some time to complete a pre-questionnaire. Before the actual test a training session was performed to familiarize the test subjects with the procedure and the range of qualities. There were four training videos containing short pieces from two of the full length videos (S and F). The training video did not contain any audio. The test leader was available during training session and answered questions if something still was unclear. Then the main part of the experiment commenced.

The main part of the experiments contained the eleven 6 minute long videos (cf. TS in Figure 1) with audio, divided into two sessions with a break after five or six video, depending on the play list. There were four different playlists prepared with different playing orders of the videos. They were not completely random. The same content (variant 1 or 2) was not played in the same session. Two playlists had six videos before the break and the other two had five videos before the break. In the break the test subjects were encouraged to go out of the Lab and possibly have some refreshments.

After the main part a post-questionnaire was filled in and then the experiment was finished with that the test subjects were thanked for their participation with a cinema ticket and a cinema gift card of 100 SEK.

Objective Characterization

In order to identify the video (PVS) characteristics influential on QoE of adaptation, different NR metrics that measures different types of video artifacts, such as blockiness, blurring, and noise, were calculated. Since the NR measures are frame based, different temporal pooling techniques such as averaging, standard deviation, Minkowski, and weighted average are tested. The results subsequently were compared to the MOS results from our previous study [11] (cf. section Subjective Experiment).

For comparison and to better understand the dataset we also consider two state-of-the-art FR VQA methods, which are all presented in the following.

No-Reference Objective Characterization Tools

To measure different video characteristics in a NR setting we rely on a selected subset of the algorithms described in [11], [13]-[14] that are publicly available at [15]. The flickering metric from [16] was also considered initially, but due to low performance it was discarded for the final results. Additionally, we also consider how well the bitrate can be used for objective characterization. The chosen measurements are briefly outlined in this Section. All the algorithms output a single measurement per frame, except the Temporal Activity (TA) measure that outputs a single measurement for each consecutive pair of frames.

Blockiness: The metric to measure the blocking artifact is calculated for pixels at boundaries of N×N blocks. It is based on the magnitude of Luma differences at the block's boundary and the picture contrast near boundaries. Since the size of the Macro Blocks (MB) in our dataset is 16×16, this is also the maximum size of the transform blocks. We therefore set N=16 when calculating the blockiness with this approach.

Blur: The metric to measure blur is based on the width of sharp edges in the image. The edges are produced by Sobel-filtering and the amount of blur is defined as the average width of the resulting edges.

Brightness: The brightness metric is based on the mean value of the average luminance level in the brightest and darkest blocks of a grayscale image.

Contrast: The metric to measure contrast is based on the assumption that the visible contrast is related to the so-called root-mean-square contrast defined in [17].

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

Noise: The metric to measure noise is based on calculations of local variance in areas with low spatial complexity.

Bitrate: The bitrate is not a metric, but is also used for objective characterization in this work, due to the relation between bitrate and video quality. Like the other measurements it is calculated per frame, so it can be calculated over the whole video with different temporal poolings.

Spatial Activity: The Spatial Activity (SA) measure is based on the spatial perceptual information measure from [18].

Temporal Activity: The Temporal Activity (TA) measure is based on the temporal perceptual information measure from [18].

Full-Reference Tools

For comparison to the NR tools and in order to better understand the dataset we also consider a state-of-the-art FR VQA method as briefly described in this Section. Besides the VQA method described below, the Visual Difference Predictor (HDR-VDP-2) presented in [19] in low dynamic range mode was tested. However, due to poor results it was excluded in the initial testing phase.

The Video Quality Model with Variable Frame Delays (VQM-VFD) presented in [20] is an improved version of the Video Quality Model (VQM) [21], but unlike VQM it does not include color parameters and it belongs to the FR category of quality assessment. The VQM-VFD model is otherwise partly based on parameters similar to those of VQM and partly based on new parameters. In VQM-VFD a neural network is used to map the values of the total eight parameters to an overall measure of distortion. The parameters are presented along with a brief description of the kind of distortion they measure:

- si_loss: Blurring.
- hv_loss: A shift of edges from horizontal/vertical orientation to diagonal orientation.
- hv_gain: Tiling or blocking.
- si_gain: Edge sharpening.
- ti_gain: Transient distortions.
- RMSE_gain: Root MSE (RMSE) in space-time blocks.
- VFD_Par1: Frame freezing.
- VFD_Par1 × PSNR_VFD: The product of temporal and spatial distortions.

Due to the good performance of the NR blockiness measure in our initial test phase, the hv_gain parameter from VQM_VFD, which is a FR blockiness measure was also extracted and used as a measure in the results.

Temporal Pooling

Since the NR objective characterization measures are frame based, techniques for temporal pooling are presented in this Section. In total 10 temporal pooling techniques are tested.

A simple temporal pooling is calculating the average:

$$\mu = \frac{1}{n_f}\sum_{i=1}^{n_f} m_i \tag{1}$$

where n_f is the total number of frames and m_i is the value of a measure for frame i.

Since frames with low perceptual quality can influence the overall quality perception of a video clip more than the rest of the frames, we also calculate the average temporal pooling with a subset of frames F_l that corresponds to 10% of the frames with the lowest measured values:

$$\mu_l = \frac{1}{n_l}\sum_{j=1}^{n_l} \breve{m}_j \tag{2}$$

where n_l is the total number of frames in the subset F_l and $\breve{m}_j$ is the value of a measure for frame j in the subset.

Since it is not clear for all measures what the influence factor of the values on the quality perception is, i.e. whether they can be regarded as a measure of distortion or of quality, a similar measure as (2) but with a subset of frames F_h that corresponds to 10% of the frames with the highest measured values is also calculated.

As a simple way to test the impact of part of the temporal aspect on the quality perception, such as the forgiveness effect, three other temporal pooling algorithms where the average of a subset is calculated similar to (2) are computed. The subsets are the frames corresponding to: the start of the video (first 2 seconds) F_s, the end of the video (last 2 seconds) F_e, and the union of those two sets $F_{se} = F_s \cup F_e$.

Another simple temporal pooling is given by the standard deviation:

$$\sigma = \sqrt{\frac{1}{n_f}\sum_{i=1}^{n_f}(\mu - m_i)^2} \tag{3}$$

This temporal pooling calculation is sensitive to many and/or large variations in the measured quality values.

Minkowski summation as detailed in [22] can also be used for temporal pooling. This technique will put emphasize on larger values if the value of Minkowski exponent parameter is higher. In this work the Minkowski summation is used with the Minkowski exponent fixed to 2 giving a little higher importance to larger values:

$$Minkowski = \sqrt{\frac{1}{n_f}\sum_{i=1}^{n_f} {m_i}^2} \tag{4}$$

An asymmetrical temporal pooling is introduced by Ninassi et al. in [23] that consists of adding the average of the measured value and a term representing the variation over time that favors the distortion decrease (compared to distortion increase).

Finally, due to the sensitivity of the human visual system to changes in quality and the forgiveness effect we also calculate a weighted average:

$$\mu_w = \sum_{i=1}^{n_f} w_i m_i \tag{5}$$

where w_i is the weight for frame i. w_i is given by a modified cosine function:

$$w_i = \mathrm{norm}\left[\cos\left(\frac{2i\pi}{n_f}\right) + \frac{i}{n_f} + 2 + s_i\right] \tag{6}$$

where $\mathrm{norm}[\cdot]$ is the function of normalization such that $\sum_{i=1}^{n_f} w_i = 1$ and s_i are local cosine waves placed in alignment

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

with the bitrate steps, such that more weight is put on measurement values close to a bitrate step. An example for w_i is given in Figure 2.

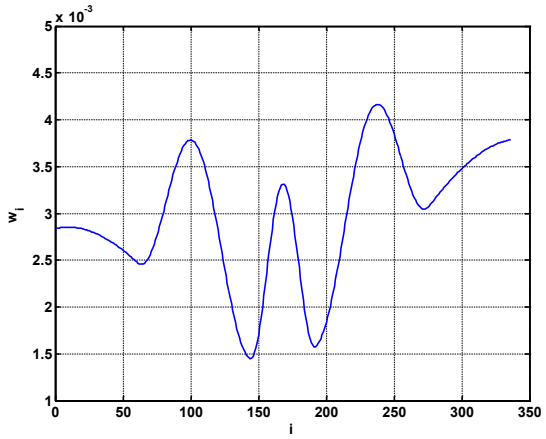

Figure 2. An example of the weighting function w_i from (6) for a sequence with 336 frames and in total three bitrate steps at frame locations: 100, 168, and 238.

Results

In this section we present our results using the methodologies outlined in the previous Sections for the subjective analysis and the objective characterization.

Subjective Analysis

The results from the Acreo subjective test was compared with the overall opinion scores collected at the earlier experiment at UPM. Figure 3 shows the MOS from the Acreo study for TS (cf. Figure 1) with 95% confidence intervals (blue bars) and the overall MOS from the UPM experiment (red bars) when the audio was presented to the subjects as well. An independent two-sample Student T-test, compensating for multiple comparison using Bonferroni correction [24] by setting alpha to 0.05/11 = 0.0045, gives the results' comparison in F_V1 and R to be statistically significantly different with p = 0.0040 and p = 0.00037.

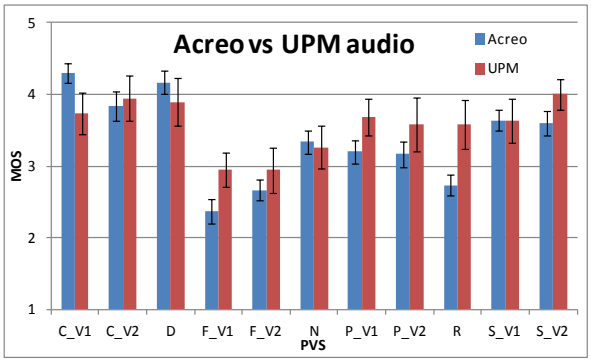

Figure 3: Comparison between the MOS obtained at the experiment conducted at Acreo and the overall scores collected at UPM containing audio. Error bars illustrates 95% confidence intervals.

Figure 4 shows Acreo's MOS scores compared to the UPM no-audio MOS. It can be noted that difference between the results is less in this case and there was no statistically significant difference. This is also illustrated with the scatter plots shown in Figure 5, where the Acreo MOS values are compared to the overall UPM audio MOS to the left and the no-audio MOS to the right. The Acreo MOS values are plotted along the x-axis and the MOS at UPM along y-axis. The correlation values are also clearly different between these two cases with 0.79 for Acreo and UPM audio and 0.93 for Acreo and UPM no-audio.

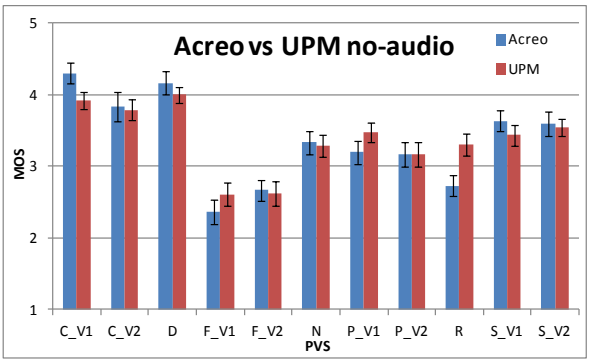

Figure 4: Comparison between the MOS obtained at the experiment conducted at Acreo and the overall scores collected at UPM containing no-audio. Error bars illustrates 95% confidence intervals.

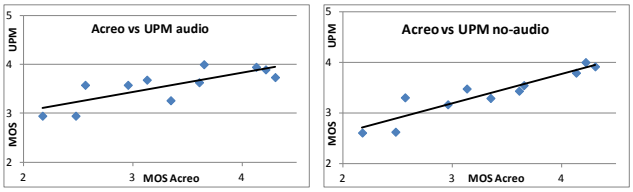

Figure 5: Scatter plots with trend lines for the MOS at Acreo compared to UPM audio MOS in graph to the left and no-audio to in the graph to the right. Acreo MOS are plotted along the x-axis and the MOS at UPM along y-axis.

One of the purposes of doing the experiment at Acreo was to compare different ways to aggregate the MOS for individual adaptation events with the overall impression MOS. In Table 4 the Pearson correlation are shown between the Acreo MOS and the overall MOS, mean MOS of last 5 adaption events (*Last 5*), MOS of last (*Last*) and the mean MOS of all adaption events (*Mean*) for both the audio and the no-audio case. Note, the high values for the mean. This is inline with the earlier finding when comparing these values in the UPM experiment, where we got 0.97 for the audio case and 0.99 for the no-audio case [11].

Table 4: Pearson correlation between Acreo overall MOS and UPM overall and aggregated MOS.

Pearson Correlation Acreo and UPM							
Audio				No-Audio			
Overall	Last 5	Last	Mean	Overall	Last 5	Last	Mean
0,79	0,66	0,54	0,81	0,93	0,70	0,71	0,90

Objective Characterization

Initially, it was found that three sequences from the dataset had low MOS scores compared to similar content and quality levels. One of these sequences mostly consisted of content that can be regarded as advertisement. The two others had abrupt scene

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

changes within the first few frames, such that the beginning of the sequences looks much more distorted than they are. Therefore, these three sequences were excluded from the rest of the experiments.

A heatmap showing the Spearman Rank Order Correlation Coefficient (SROCC) between measurements and the MOS with different temporal pooling techniques is shown in Figure 6. In general, it appears that the correlation is very low. The best correlation obtained is 0.46 and -0.46 for the blockiness measure with the average pooling of the 10% lowest values and the standard deviation, respectively. The brightness, noise, and TA measures have negative correlation with the MOS, while contrast and bitrate have positive correlation with the MOS for most pooling methods. Blur and SA seems to be rather inconsistent and generally only low correlation with MOS were obtained for these measures. In comparison the correlation obtained by the FR metric VQM_VFD and by the FR measure hv_gain were 0.68 and 0.44, respectively.

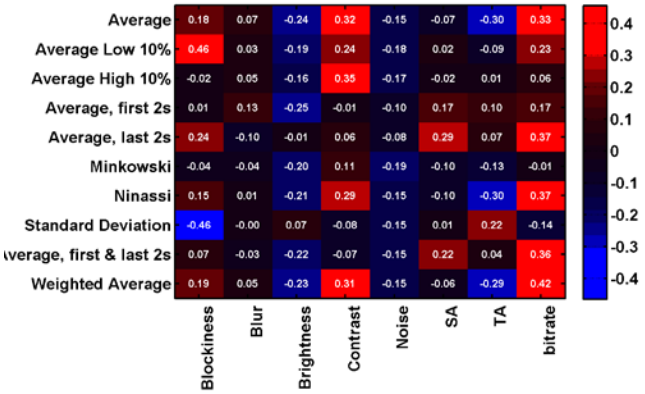

Figure 6. SROCC values between NR objective measurements and the MOS using the selected temporal pooling techniques.

The different content in the dataset has different characteristics and therefore the SROCC was also calculated for subsets of sequences categorized by content. The results can be seen in Figures 7 and 8 with average pooling and average pooling of the lowest 10% of the values for the NR measurements, respectively. As expected, this clearly improves the correlation with the MOS for most measurements and subsets. It can also be seen that quality prediction seems to be especially difficult for two subsets, the two versions of the Football content. This content generally has high temporal complexity and for one version the TA measure is actually the measurement with best SROCC to the MOS. The SROCC for the NR and FR blockiness measure is also lower for this content, which is limits the effectiveness of VQM_VFD. On the other hand, the noise measure has higher SROCC values for this content. Interestingly, in some cases there also seems to disparity in the SROCC between measurements and MOS for the two versions of the same original content. This is also pronounced for the Football content.

The oddities with the Football content could be due to fact that the content is a mix of e.g. scenes with the football game itself with high spatial and temporal complexity, scenes showing the audience of the match, and scenes with lower complexity showing trainers and players. The disparity in the performance of the measures between the two versions of the Football content, seems to be due that the type of scenes described above has very different impact on the video coding, therefore leading to less or more artifacts in the different versions.

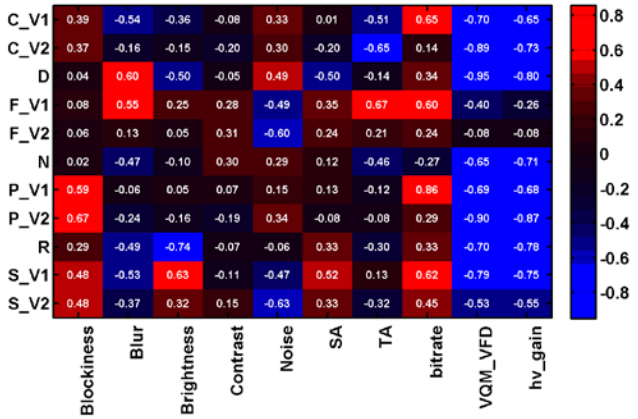

Figure 7. SROCC values between NR and FR measurements divided into subsets of content. Average pooling was used for the NR measurements.

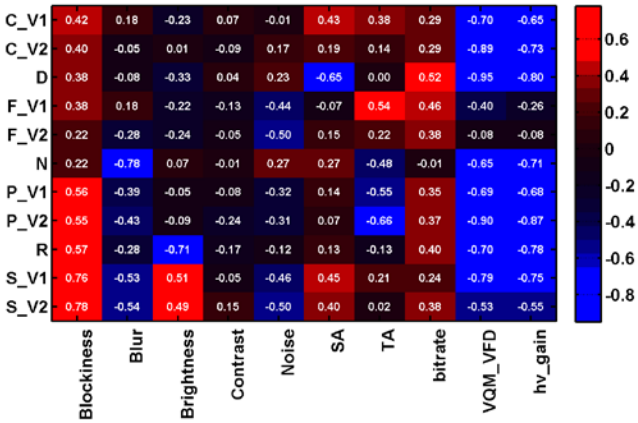

Figure 8. SROCC values between NR and FR measurements divided into subsets of content. Average pooling of the lowest 10% of the values was used for the NR measurements.

As a step towards using objective characterization to improve quality prediction, we performed clustering in the space of spatial and temporal complexity as expressed by the SA and TA values. The result of this is shown in Figure 9. Each cluster includes several different content. The obtained SROCC values in each of the clusters were 0.42, 0.57, 0.82, and 0.71 for average pooling of the lowest 10% of the NR blockiness values. All of these values, except one (cluster 1), are much higher than without clustering. For the one exception, other measures than blockiness might be more relevant for this part of the SA/TA space. Similar results were obtained for the SROCC in clusters for the FR blockiness measure.

The results presented in this Section are a selection of the findings with objective characterization for adaptive streaming videos. Other results that might also be relevant, but need further investigation, such as the impact of scene changes on the perceived quality, is left for future work.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-105

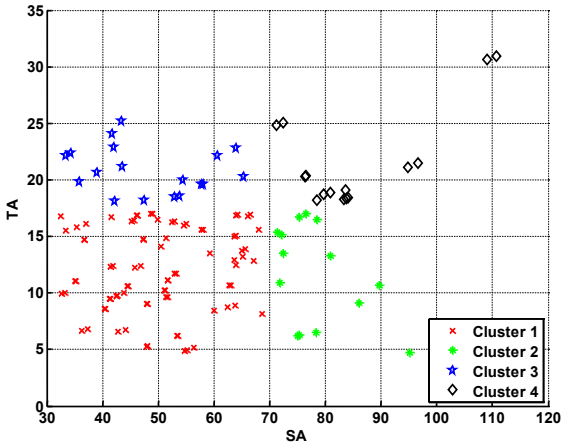

Figure 9. Clustering of the dataset into four clusters based on the SA and TA values.

Discussion

The subjective experiment conducted at Acreo showed some interesting differences with earlier results conducted at UPM. One of the reasons for doing the experiment at Acreo was to investigate whether the test subjects at UPM when giving the overall scores were affected by rating the individual events. The differences found could to some extent be attributed to this fact. However, there were other important differences as well such as language and cultural. We chose to present exactly the same videos at Acreo as were used in the previous experiment at UPM. The sound track contained therefore both Spanish and English language. Especially at the Spanish language videos we found big differences (P and F). The test panel in Sweden did understand the English soundtrack but in most cases not the Spanish. The test panel in Spain understood probably the Spanish soundtrack better than the English. Exceptions that could not be explained by this was N (news content) which was in Spanish (very little difference) and R (music concert) which was in English (biggest difference).

If we compare the overall scores collected at Acreo with the no-audio scores at UPM, then there is a clearly higher similarity. It seems like the test panel at Acreo focused more on the visual quality despite that the audio was present. This could have been reinforced through the fact that some of the videos had unfamiliar language emphasizing that the visual aspect was the main target. In the post-questionnaire question about if the test subjects thought that the language has had an impact on their scoring 86 % answered that it had not, confirming that they had focused on the visual quality.

We investigated also different methods to aggregate the individual scores to an overall rating. The tested methods were to take the MOS of the last adaptation event, the mean of the MOS of the last five adaptation events or the mean of all adaptation events. From the earlier experiments we found an almost perfect correlation (0.94 for audio and 0.99 no-audio) between the mean MOS and the overall impression, which could be somewhat attributed to that both the individual scores and the overall scores had been given by the same test subjects at the same occasions. We still find that the mean of MOS of the individual events had a good correlation if we compare to MOS obtained at Acreo, although not as high (0.81 for audio and 0.9 for no-audio), confirming the earlier result.

On the other hand, due to the difficulty of quality prediction of adaptive streaming videos with realistic content, objective characterization could very well be vital to improve future video quality assessment. In this respect we showed a selection of tools that can be used for objective characterization of video content. Machine learning techniques can be employed to find the contribution of each characteristic on quality perception.

We also found that the performance of employed tools depends on the video type. There was specialy quite low performance when on the content of a football match (F) due to the characteristics of the original content.

Conclusions

The main finding of this paper on the subjective results is that mean of the MOS of the individual adaption events is a good predictor of the overall MOS for the full length 6 minute video sequence.

The objective characterization of adaptive streaming video in our data set had in general low performance with the subjective scores. However, the performance could be increased by clustering on SA and TA. The FR blockiness was the best predictor especially if it were aided.

Acknowledgement

The help preparing and conducting the experiment at Acreo by Börje Andrén, Anders Djupsjöbacka, Jesus Gutierrez and Kun Wang as well as the economic support from VINNOVA (Sweden's innovation agency), EIT Digital, Ministerio de Economa y Competitividad of the Spanish Government (project TEC2010-20412) are hereby gratefully acknowledged.

References

[1] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake, "Quality of Experience and HTTP Adaptive Streaming: a Review of Subjective Studies," in Proc. Of IEEE 6th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 141–146, Sep. 2014.

[2] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hossfeld, and Tran-Gia P., "A Survey on Quality of Experience of HTTP Adaptive Streaming," in IEEE Communications Surveys & Tutorials, 2014.

[3] International Telecommunication Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU-R Recommendation BT.500-13, 2012.

[4] C. Chen, L. Choi, G. de Veciana, C. Caramanis, R. Heath, A. Bovik, "Modeling the Time Varying Subjective Quality of HTTP Video Streams with Rate Adaptations," IEEE Transactions on Image Processing, vol. 23, no. 5, pp. 2206–2221, 2014.

[5] M.H. Pinson, M. Sullivan, and A. Catellier, "A new method for immersive audiovisual subjective testing," in VPQM, 2014.

[6] S. Tavakoli, J. Gutierrez, N. Garcia, "Subjective Quality Study of Adaptive Streaming of Monoscopic and Stereoscopic Video," IEEE Journal on Selected Areas in Communications, vol. 32, no. 4, pp. 684–692, 2014.

[7] J. Lievens, A. Munteanu, D. De Vleeschauwer, and W. Van Leekwijck, "Perceptual video quality assessment in HTTP adaptive streaming," IEEE Int'l Conf. on Consumer Electronics (ICCE), pp. 72–73, 2015.

[8] K. Singh, Y. Hadjadj-Aoul, and G. Rubino, “Quality of Experience Estimation for Adaptive HTTP/TCP Video Streaming Using H.264/AVC,” in IEEE Consumer Communications and Networking Conf. (CCNC), pp. 127–131, 2012.

[9] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, “To Pool or Not to Pool: A Comparison of Temporal Pooling Methods for HTTP Adaptive Video Streaming,” in Proc. of IEEE 5th Int’l Workshop on Quality of Multimedia Experience (QoMEX), pp. 52–57, 2013.

[10] J. Søgaard, S. Forchhammer, and K. Brunnström, ”Quality Assessment of Adaptive Bitrate Videos using Image Metrics and Machine Learning”, In Proc. of IEEE 7th Int’l Workshop on Quality of Multimedia Experience (QoMEX), pp. 1-2, May 2015.

[11] S. Tavakoli, K. Brunnström, J. Gutierrez, and N. Garcia, “Quality of experience of adaptive video streaming: Investigation in service parameters and subjective quality assessment methodology,” Signal Processing: Image Communication, Special Issue on Recent Advances in Vision Modelling for Image and Video Processing, vol. 39-B, pp. 432–443, Nov. 2015.

[12] M. Leszczuk, M. Hanusiak, M. Farias, E. Wyckens, G. Heston, “Recent Developments in Visual Quality Monitoring by Key Performance Indicators”, Springer Multimedia Tools and Applications, pp. 1-23, Sept 2014.

[13] M. Leszczuk, M. Hanusiak, I. Blanco, A. Dziech, J. Derkacz, E. Wyckens, S. Borer, “Key Indicators for Monitoring of Audiovisual Quality”, Signal Processing and Communications Applications Conference (SIU), 2014.

[14] M. Mu, P. Romaniak, A. Mauthe, M. Leszczuk, L. Janowski, E. Cerqueira. "Framework for the Integrated Video Quality Assessment." Multimedia Tools and Applications 61, no. 3, pp. 787-817, 2012.

[15] Department of Telecommunications, AGH University of Science and Technology, “Video Quality Metrics”, URL: http://vq.kt.agh.edu.pl/metrics.html

[16] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, “Perceptual Quality Assessment for H.264/AVC Compression,” in IEEE Consumer Communications and Networking Conference (CCNC), pp. 597–602, 2012.

[17] E. Peli, "Contrast in Complex Images," Journal Optical Society of America A 7, pp. 2032-2040, 1990.

[18] International Telecommunication Union, “Subjective Video Quality Assessment Methods for Multimedia Applications,” in ITU-T Rec. P.910, 2008.

[19] R. Mantiuk, K. Kim, A. Rempel, and W. Heidrich, “HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions“, ACM Transactions on Graphics (Proc. of SIGGRAPH), vol. 30, no. 4, no. 40, 2011.

[20] M. Pinson, L. Choi, and A. Bovik, “Temporal Video Quality Model Accounting for Variable Frame Delay Distortions,” IEEE Trans. on Broadcasting, vol. 60, no. 4, pp. 637–649, Dec 2014.

[21] M. Pinson and S. Wolf, “A New Standardized Method for Objectively Measuring Video Quality,” IEEE Trans. On Broadcasting, vol. 50, no. 3, pp. 312–322, 2004.

[22] S. Rimac-Drlje, M. Vranjes, and D. Zagar, “Influence of Temporal Pooling Method on the Objective Video Quality Evaluation,” IEEE Int’l Symposium on Broadband Multimedia Systems and Broadcasting, 2009.

[23] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, “Considering Temporal Variations of Spatial Visual Distortions in Video Quality Assessment,” IEEE Journal Selected Topics in Signal Processing, 2009.

[24] K. Brunnström, S. Tavakoli, and J. Søgaard, “Compensating for Type-I Errors in Video Quality Assessment,” In Proc. of IEEE 7th Int’l Workshop on Quality of Multimedia Experience (QoMEX), pp. 1-2, May 2015.

Author Biography

Jacob Søgaard received the B.S. degree in engineering, in 2010, and the M.S. degree in engineering, in 2012, from the Technical University of Denmark, Lyngby, where he is currently pursuing his Ph.D. degree with the Coding and Visual Communication group at the Department of Photonics. His research interests include image and video coding, image and video quality assessment, visual communication, and machine learning in the context of Quality of Experience.

Samira Tavakoli received the Master degree in Telecommunication Engineering from the Blekinge Tekniska Högoskola (BTH), Sweden, in 2010. In 2015, she finished her Ph.D. thesis on "Subjective QoE Analysis of HTTP Adaptive Streaming Applications" in the Universidad Politécnica de Madrid (UPM), Spain. Since 2010, she is a member of the Image Processing Group (GTI) at the UPM. From 2012, she has been working with Acreo Swedish ICT AB in the area of subjective quality studies. Her research interests include on evaluation of user-oriented techniques in the ICT domain and advances in laboratory methodologies.

Kjell Brunnström, Ph.D., is a Senior Scientist at Acreo Swedish ICT AB and Adjunct Professor at Mid Sweden University. He is an expert in image processing, computer vision, image and video quality assessment having worked in the area for more than 25 years. Currently, he is leading standardization activities for video quality measurements as Co-chair of the Video Quality Experts Group (VQEG). His current research interests are in Quality of Experience for visual media in particular video quality assessment both for 2D and 3D, as well as display quality related to the TCO requirements.

Narciso García received the Doctor Ingeniero de Telecomunicacíon degree (Ph.D. in Communications) in 1983 (Doctoral Graduation Award) from the Universidad Politécnica de Madrid (UPM), Madrid, Spain. Since 1977 he has been a member of the faculty of the UPM, where he is currently a Professor of Signal Theory and Communications. He leads the Grupo de Tratamiento de Imágenes of the UPM. He has been actively involved in Spanish and European research projects, serving also as evaluator, reviewer, auditor, and observer of several research and development programs of the European Union. He was a co-writer of the EBU proposal, base of the ITU standard for digital transmission of TV at 34-45 Mb/s (ITU-T J.81). He was Area Coordinator of the Spanish Evaluation Agency (ANEP) from 1990 to 1992 and General Coordinator of the Spanish Commission for the Evaluation of the Research Activity (CNEAI) since 2011-2014. He was awarded the Junior and Senior Research Awards of the Universidad Politécnica de Madrid in 1987 and 1994, respectively. His professional and research interests are in the areas of digital image and video compression and of computer vision.