Impact of HEVC encoding presets for resolutions up to UHD-1

Christopher Krämmer
Institute for Media Technology
TU Ilmenau
Email: christopher.kraemmer@tu-ilmenau.de

Serge Molina
Systèmes Robotiques et Interactifs
UPSSITECH
Email: serge.molina@kloumpt.net

Anton Schubert
Institute for Media Technology
TU Ilmenau
Email: anton.schubert@tu-ilmenau.de

***Abstract*—TODO: shorten**

Online video accounts for a large amount of the Internet traffic. To adapt better to changing bandwidth requirements more and more content distributors employ HTTP Adaptive Streaming (HAS), however specific quality measures are not yet generally available. This paper aims at linking predicted quality and subjective user ratings based on the *VMAF* metric and an *ACR*-style experiment. We analyze encodings of six different 10-second sequences at varying resolutions up to *UHD-1*, with three bitrates per resolution, and compare them with ratings from the perceptual test based on two different encoding presets (a "naïve" approach and an "expert" one). We could accurately predict our final subjective quality results using the *VMAF* metric. Our aggressive optimization with the "expert" encoding preset saves bits and preserves quality at higher bitrates, but leads to visual quality degradation for low bitrates.

***Index Terms*—UHD, videoquality, VMAF**

1. Introduction

Broader access to online media content, especially on mobile devices [1], combined with an growing diversity of content has led to an increased saturation of internet backbone networks [2]. More dynamic delivery approaches like HTTP Adaptive Streaming (*HAS*) [3] are applied to increase reliability for changing bandwidth environments, while still maintaining a high quality of experience for the consumer. However, this requires that the media content is preprocessed before delivery with video encoding and segmentation. The optimum choice of encoding parameters depends on a number of factors and is often hard to determine as generalized quality models for Adaptive Streaming are still in development [4] and not yet generally available.

This paper aims at finding a link between predicted quality and user ratings based on the Video Multi-Method Assessment Fusion (*VMAF*) metric [5], [6] and a subjective quality experiment following ITU P.910 recommendations [7]. We analyze encodings of six different 10-second sequences at varying resolutions up to *UHD-1* [8], with three bitrates per resolution, and compare them with ratings from the perceptual test based on two different encoding presets (a "naïve" approach and an "expert" one).

We describe our selection of the source sequences, the encoding process and the setup for the perceptual test. As achieving reproducible results is a strong focus of this research project the preprocessing and video encoding steps have been largely automated.

Several data has been collected in addition to user ratings, such as feedback questions and rating behavior. This additional data has been collected in order to enable a deeper analysis of the subjective test results.

2. Test Preparation

This section will show the different steps followed during the preparation of the subjective experiment. These steps will be described as precisely as possible in order to guarantee reproducibility of our results. At first the video selection methodology and the associated sources are introduced in section 2.1. Following that, section 2.2 describes how the source sequences are encoded, and finally section 2.3 presents the test setup with its dedicated rating framework.

2.1. Video Selection

The selection of suitable sequences for the subject test is based on technical considerations and content.

2.1.1. Preconsiderations. The number of public video databases with royalty free *UHD-1* and 4k video material is limited. They differ in content (scenario, field size, camera movement, illumination condition) and in technical aspects (resolution, frame rate, bitrate and color bit depth). To obtain reference video files for the subjective test, parts of the *UHD-1* content from the databases are used (also called source video files). For our test we want to have a large variety between the reference video files to evoke different encoding properties. All the reference videos should have a duration of 10 seconds as they represent a single *HAS*-segment and should contain no hard cuts. To avoid the influence of judder, the smallest permitted frame rate of the source videos is 50 frames per second (fps). Furthermore the smallest resolution considered to be 3840x2160. Moreover the frame rate of the reference video files is being adapted to consistent 50 fps and the resolution to 3840x2160 pixels.

2.1.2. Dataset Preparation. For obtaining a good variety of video sequences three databases are used: The Harmonic [9] which contains 18 different video files, Cable Labs [10] with 9 relevant *UHD-1* contents and the Blender Foundation [11] for receiving the cartoon Big Buck Bunny. They are partial under the creative commence license and all of them are available in the *ProRes* format, except of Big Buck Bunny where the *UHD-1* video content can be downloaded only as a compressed video file, however there also exist high quality PNGs for all the frames of the cartoon. In order to generate a high quality reference video with the frames of Big Buck Bunny, an automatisation script is used to find a sequence with a duration of 10 seconds, to download the respective images and to encode them as a video file. The main challenge is finding periods of video without cuts that are 10 seconds or longer, as only a few of those are available, which is also a limiting factor for the content diversity.

One problem of the video sequences from the database is that even if they have high bitrates and are stored as *ProRes*, no assumptions can be made on the visual quality of the material. It would be preferred to have the source videos available in RAW format as this assures no loss of information and no previous processing. For this reason a preselection with a system able to play UHD content in *ProRes* with high bitrates and connected to a 4k screen was applied. The remaining reference videos are 6 source clips, each with a duration of 10 seconds. All the files exhibit a minimum resolution of 3840x2160 pixels and are stored in *ProRes*. Because this format is a 10 bit codec only [12], the original color depth of each video file can not be determined. Further specific information about the reference videos are shown in Table 1, e.g. Air Show, a video from the Harmonic Database, available with 59.94 fps and a bitrate of 1703 Mbit/s. For obtaining the reference video with a duration of 10 seconds the extraction starts at the 48.5 second of the source video file.

TABLE 1. META DATA OF THE VIDEO FILES. THE TIMESTAMPS ARE THE START POSITIONS OF THE EXTRACTION OF 10 SECOND SEQUENCES FROM THE SOURCE FILES, WHERE MM STANDS FOR MINUTES, SS FOR SECONDS AND MS FOR MILLISECONDS RESPECTIVELY. FURTHER SHORTCUTS: H = HARMONIC, B = BLENDER FOUNDATION, C = CABLE LABS

Sequence Name	Frame Rate fps	Bit rate Mbits/s	Timestamps mm:ss.ms	Source
Air Show	59.94	1703	00:48.500	H
Big Buck Bunny	60	2304	05:47.000	B
Fjord	50	1469	00:21.000	H
Moment of Intensity	59.94	1822	02:16.000	C
Snow Monkeys	59.94	1750	00:17.000	H
Streets of India	50	2094	00:00.000	H

They contain a wide range of high-level features (animation, camera motion, people, water) and low-level characteristics (brightness, contrast, texture, motion, color variance, sharpness) as can be seen in Figure 1.

(a) Air Show (b) Big Buck Bunny (c) Fjord (d) Moment of Intensity (e) Snow Monkeys (f) Streets of India

Figure 1. Selection of frames contained in the reference videos to show the variety of the content.

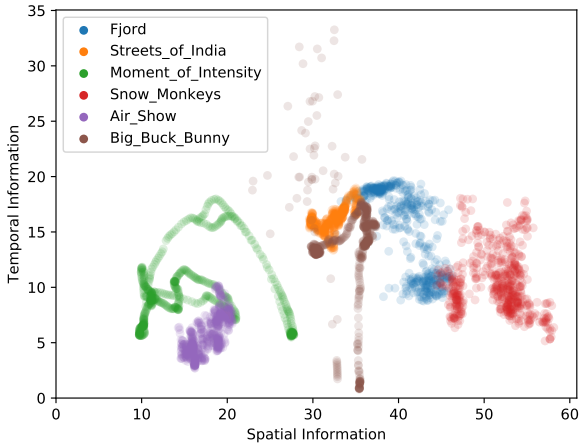

Figure 2. Spatial and temporal information of the reference video files. Each dot represents one frame. In areas with high densities the color is stronger.

We further analyzed the spatial and temporal changes between the frames of each reference video with SITI [13], a command-line-based tool that refers to ITU-T P.910. The results can be seen in Figure 2. The spatial information is derived as the maximum standard derivation of the results of a sobel filter applied on each frame. Furthermore the temporal information is the maximum standard derivation of the motion difference between two frames at the same location in space. As expected, the Air Show sequence includes the smallest spatial and temporal changes in comparison with the other reference videos due to the only slightly changing,

low complexity scenario. In contrast to that the Big Buck Bunny sequence shows a large range of temporal differences because of camera movement, while the Snow Monkeys sequence has the highest spatial information due to fine structures and details.

The following section goes into detail on how the chosen sequences are encoded for the subjective test.

2.2. Encoding Parameters

As one of the focus points of this paper the choice of encoding parameters and the video preprocessing are of main importance. This section will first cover the two different encoding presets, go over the method which is used to derive the target bitrates and finalize with a description of the process automation.

2.2.1. Encoding Presets. We use the open x265 encoder for our experiment as it offers good performance and integration in the FFmpeg toolchain. Two different presets are used for the sequence encodings, a ”naïve” (1) and an ”expert” preset (2). The ”naïve” preset is a simple *CBR* (Constant Bitrate) encoding, whereas the ”expert” preset is a 2-pass encoding with a Quality-Control pass followed by a Bitrate-Control pass. Every sequence is encoded with both presets at 3 resolutions (540p, 1080p, 2160p) and 3 bitrates for each resolution. The following section goes into detail on how we select those bitrates.

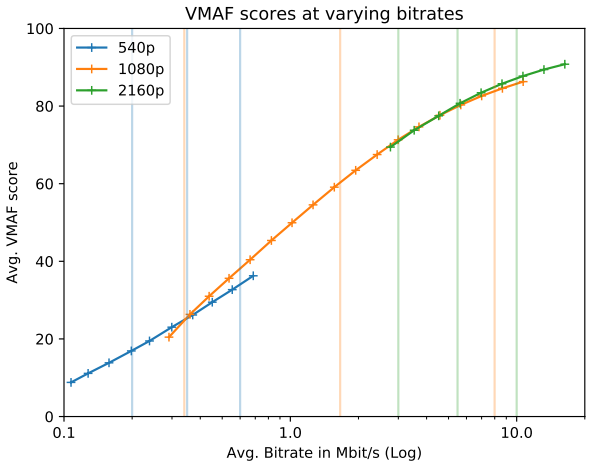

Figure 3. Average *VMAF* scores for 25 different bitrates at 3 resolutions. The encoded bitrates and the *VMAF* scores are averaged between the 6 sequences and form the abscissa (Log) and ordinate respectively.

2.2.2. Selection of Bitrates. Video Multi-Method Assessment Fusion (*VMAF*) is a full reference metric for estimating human perception of video quality [5]. We use *VMAF* because it provides a better estimate of subjective quality than single metrics like *SSIM* or *VIF*.

To estimate relevant *HEVC* encoding bitrates for our source content we sample the *VMAF* scores at 25 bitrates on a logarithmic scale for our 3 different resolutions (540p, 1080p, 2160p). The reference sequences are resampled to a fixed 50 frames per seconds to avoid frame rate differences, while the distorted sequences are downsampled, encoded with *CBR* rate control and upsampled to *UHD-1* again using lanczos resampling. Both presets use 4:2:0 chroma subsampling to be close to the typical use-case of webvideo. The sampling requires 222 total sequences to be encoded with x265 and analyzed with the *VMAF Development Kit* (VDK). This process takes around 18 hours (excluding download and cutting of the source material) on a current 10-Core x64 CPU.

The resulting *VMAF* scores exhibit an overlap between different resolutions and the final encoding bitrates are chosen near those intersections. Figure 3 shows the sampled scores with the target rates in the background. We choose the target bitrates at least 2 bitrate-samples away from an intersection with the next quality, except for the lower 1080p-bound where it is not possible to lower the bitrate any further due to encoder restrictions. This should ensure a relevant sampling of the *MOS*-bitrate space for the subjective test.

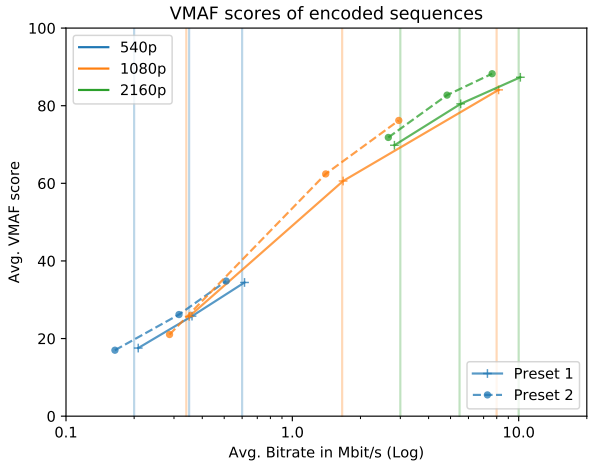

Figure 4. Average *VMAF* scores of encoded videos for both presets. The encoded bitrates and the *VMAF* scores are averaged between the 6 sequences for each preset and form the abscissa (Log) and ordinate respectively.

The average *VMAF* scores of the two presets at the chosen bitrates can be seen in Figure 4. The respective target bitrates can be seen in the background again. The scores of the ”expert” preset (2) are consistently higher than the ones for the ”naïve” preset (1) at matching target bitrates. Furthermore, the ”expert” preset saves bitrate by resorting to an acceptable level of quality, while the ”naïve” presets bitrates are very close to the targets. However, we can see that the preset two sometimes reduces the bitrate too far so that the average score is well below preset one. This happens especially for the lowest 1080p bitrate and less so for the 540p encodings.

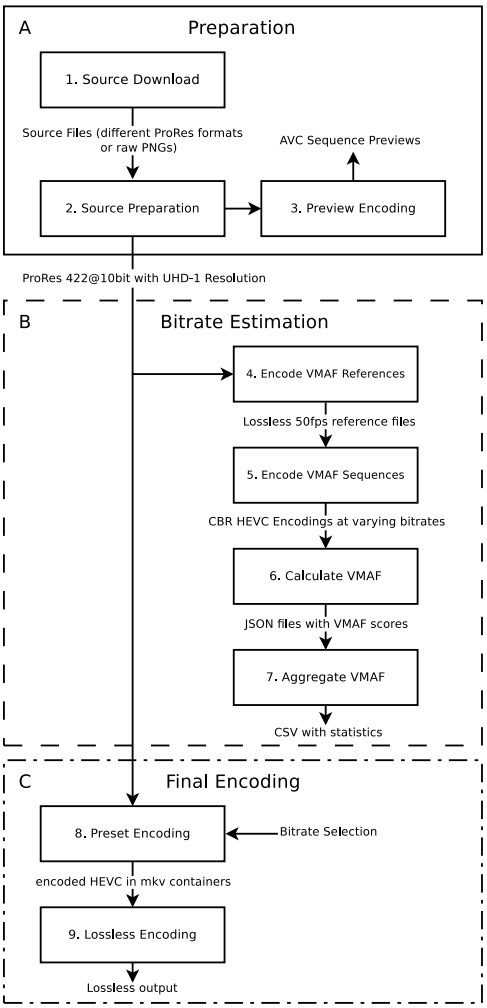

Figure 5. Automated processing and encoding workflow.

2.2.3. Encoding Automation. We automate the whole process for downloading, preprocessing and encoding the source videos using pydoit [14]. This speeds up the turnaround time for changed parameters or sequences and also ensures that the encoded material can later be reproduced.

The whole process is illustrated in Figure 5 and starts with the source preparation (A). After download of the sequences (1) they are cut to 10 seconds length and saved as ProRes HQ with *UHD-1* resolution (2). Additionally, *MPEG4-AVC* previews are generated at a lower resolution of 1440p to allow review of the sequences on slower devices.

After the initial processing the bitrate estimation is performed (B) using the *VMAF* metric [5]. The videos are brought to the same frame rate of 50fps (4) and encoded with *CBR* at 25 different bitrates (5). The average *VMAF* score of each video is analyzed with the *VMAF* Development Kit (VDK) [15] and saved as a json file (6). All of the average scores are then aggregated into a single CSV for plotting and further analysis (7). The target bitrates can then derived from these scores.

The last step is the main encoding (C). It can only start after the target bitrates have been specified in the configuration. The sequences are encoded first with the two presets (8) and transcoded to a lossless format (ffvhuff) afterwards, to allow for fast and consistent playback as well as archiving of the video material.

2.3. Test Setup

Several steps have been followed in order to enable the acquisition of meaningful data during the subjective tests. These steps include defining the test environment, choosing a rating framework and creating additional question that may help inferring informations about the participants and their way of rating content.

2.3.1. Test Environment. In order to make the results of our research reproducible we follow the ITU P.910 [7] recommendation. The parameters include aspects such as viewing distance, peak luminance of the screen or background room illumination. Performing tests following these specifications is common at the Technical University of Ilmenau, so we used a room meeting these requirements, which was already available.

2.3.2. Rating Framework. The testing procedure which seems most suited to this case is Absolute Category Rating (ACR) [7], where different versions of an original sequence are shown to a test participant.

For each sequence the participant issues categorical ratings from any of these 5 answers: {Excellent, Good, Fair, Poor, Bad}

The steps performed by a participants during a rating session can be seen in the figure 6.

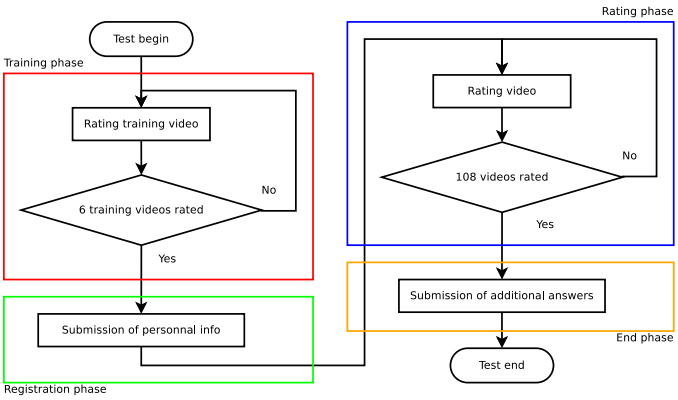

Figure 6. Detailed steps of a rating session

2.3.3. Additional collected data. In order to gain more knowledge about the users behaviour, additional data is collected from the participants.

One of the suggestions of the ITU recommendation paper being the usage of several test questions in addition

to the ratings [7], the following points have been assessed at the end of each rating session:

- Impact of the following points on rating
  - Presence of blocky artifacts
  - Visible bands of colour
  - Smoothness of the playback
- Exposure to 4k content and devices
- Confidence of the participant in his ratings

The idea behind these question is dual as they may translate how users perceive video features as well as how users may clearly express their perceptions. As it has been noticed in previous experiments and in discussions with test participants: users may still rate using a different scales, thus spreading the final *MOS* or falsely being classified as outlier when their ratings may only represent a shift from the overall population.

Moreover, mouse interactions have been collected during the rating of each sequence. The intent behind this is that as the MOS scale forbids detailed answers some participants may hesitate between two answers and change their final answer or hover over several ones before making his final choice. Answering speed, which could be an indicator of a participant skipping answers or to the contrary being strongly confident in his answers is a feature that has been thought relevant to analyse. We believe that informations such as confidence in the participants ratings or more precise scores can be extracted from these previously described behaviours.

3. Test Results and Analysis

This section focuses on presentation and analysis of the result data. At first we look at outliers in our participant data, further analyse the MOS per sequence and per preset and compare the bitrate and quality usage of the preset encodings. Finally we correlate our MOS data with the average *VMAF* scores.

3.1. MOS Ratings

In this section we look at the results of the perceptual test and analyze them.

3.1.1. Outliers. We calculated the average of all *MOS* and compared the differences of each participants ratings to the average. This can be seen in 7. All participants apart from number 12 and 17 fall into a difference range of $[-0.5, 0.5]$ to the mean. For this reason we exclude the rating data of participants 12 and 17 from our further analysis.

3.1.2. *MOS* per Sequence. In Figure 8 we show the *MOS* for each sequence aggregated over all versions of this sequence. We see that the distributions are similar over all sequences and that every sequence has ratings on all steps of the *ACR* (Absolute Category Rating) scale. This suggests a broad range of qualities in the encoded video.

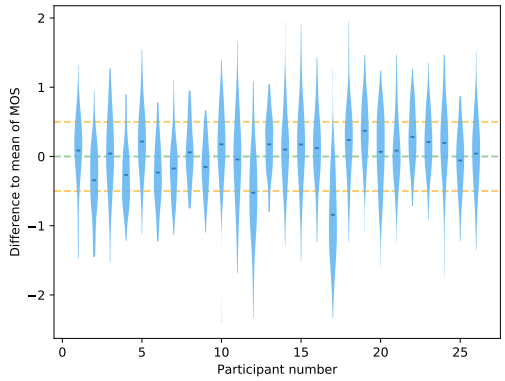

Figure 7. Difference between average MOS and the invididual ratings for each participant

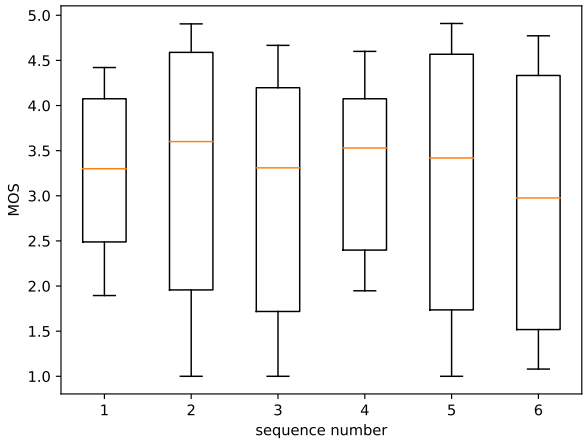

Figure 8. *MOS* distribution per sequence over all versions of the sequence - 1: Air Show, 2: Big Buck Bunny, 3: Fjord, 4: Moment of Intensity, 5: Snow Monkeys, 6: Streets of India

3.1.3. *MOS* per Preset. The Distribution of *MOS* values for each preset at different resolutions is shown in Figure 9. The ”expert” preset quality is higher at 2160p resolution compared to the other preset and also compared to 1080p. For high and medium bitrates at 1080p and 540p the quality is similar to preset 1, but the MOS values exhibit a wider spread. However, we see that especially for the lowest bitrate version the quality of ”expert” preset (2) degrades, which we could already predict from our *VMAF* scores after encoding in Fig. 4.

We further look at the bitrate differences between the presets in Figure 10. It shows the relative bitrate savings for each target bitrate. The hatched bars show where the ”expert” preset degrades the subjective quality below that of , which happens mainly for the 540p and 1080p resolutions. The 2160p versions of the content largely benefit from the ”expert” preset with high subjective quality at significantly

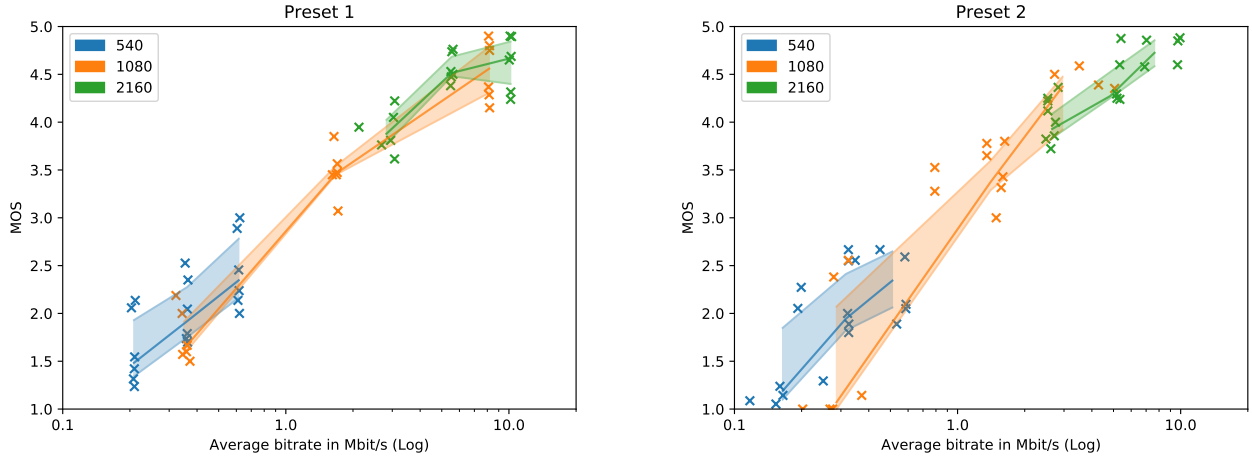

Figure 9. Correlation between bitrate and *MOS* for both encoding presets. The center line represents a median and the outer line the 25th and 75th percentile of *MOS* for the 6 sequences.

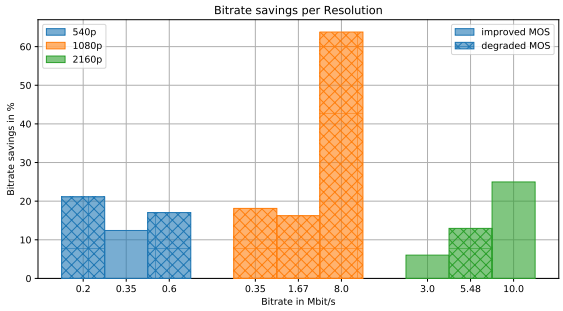

Figure 10. Relative bitrate difference of ”expert” preset compared to ”naïve” preset at all target bitrates. For target bitrates with added hatching the ”expert” preset MOS is smaller than the ”naïve” preset MOS.

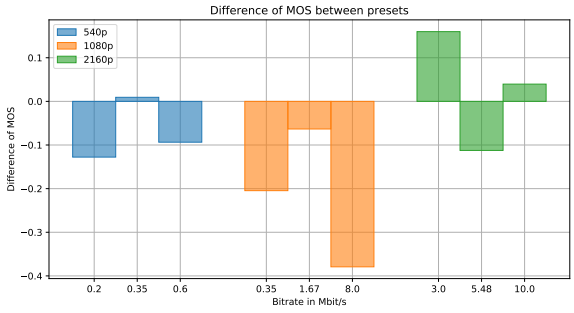

Figure 11.

lower bitrates.

The Pearson product-moment correlation between user-rating based *MOS* and the precomputed *VMAF* scores is strong at 93% (see fig. 12).

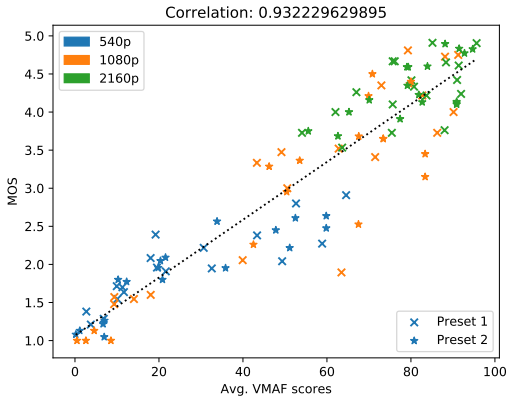

Figure 12. Correlation between average *VMAF* scores and *MOS* for each preset and resolution

3.2. Features

(Mouse tracking, Additional Questions)

4. Conclusion

4.0.1. Conclusion. The final subjective quality can be accurately predicted using the *VMAF* metric. Our aggressive optimization approach with the ”expert” encoding preset pays off at higher bitrates but degrades visual quality for low bitrates. This suggests that high-effort multi-pass encodings mainly benefit high resolution content at large bitrates, where the absolute gain can also be larger.

4.0.2. Future Work. Improvements of the automation framework: The source Preparation step could directly transcode the original material to a constant framerate and

color-subsampling in a lossless format to avoid a further preprocessing step for the VMAF metrics.

A longer test or a sectioned test with more participants would provide a better sampling of the bitrate-MOS space and allow for more detailed analysis.

The test could be done to specifically look at content-specific encoding approaches like [16].

References

[1] C. V. Forecast, "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020 white paper," *Cisco Public Information*, 2016.

[2] I. SANDVINE, "Global internet phenomena report. 2012," 2012. [Online]. Available: https://www.sandvine.com/downloads/general/global-internet-phenomena/2012/1h-2012-global-internet-phenomena-report.pdf

[3] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of http adaptive streaming," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 469–492, Firstquarter 2015.

[4] A. Raake, M. N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for http adaptive streaming: Itu-t p.1203.1," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2017, pp. 1–6.

[5] T. J. Liu, W. Lin, and C. C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1793–1807, May 2013.

[6] J. Y. Lin, T. J. Liu, E. C. H. Wu, and C. C. J. Kuo, "A fusion-based video quality assessment (fvqa) index," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–5.

[7] I. REC, "P. 910," *Subjective video quality assessment methods for multimedia applications*, 1998.

[8] "DVB UHD-1 Phase 1," *ETSI TS 101 154 V2.2.1*, pp. 124–126, 06 2015.

[9] Harmonic. (2017, Jun) Harmonic demo footage. [Online]. Available: https://www.harmonicinc.com/4k-demo-footage-download/

[10] CableLabs. 4k video database. [Online]. Available: http://4k.cablelabs.com/

[11] Big buck bunny. Blender Foundation. [Online]. Available: http://distribution.bbb3d.renderfarming.net/video/png

[12] Apple. (2017, Apr) Apple prores white paper. [Online]. Available: https://images.apple.com/final-cut-pro/docs/Apple_ProRes_White_Paper.pdf

[13] P. Lebreton and W. Robitza. (2016, Dec) Telecommunication-telemedia-assessment/siti. [Online]. Available: https://github.com/Telecommunication-Telemedia-Assessment/SITI

[14] E. Schettino. Doit automation tool. [Online]. Available: http://pydoit.org/

[15] Vmaf development kit (vdk). [Online]. Available: https://github.com/Netflix/vmaf

[16] J. D. Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 1484–1488.