Analysis of two different 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*—Abstract**

***Index Terms*—4k, videoquality, VMAF, bitrate ladder**

I. Introduction

Broader access to online content[[cite something]] combined with and increased diversity of available content[[cite something]] has lead to a saturation of internet traffic. This issue has been mitigated through changes in how content is delivered, these changes include more intelligent delivery approaches such as using content delivery networks (CDN) [1] and Adaptive BitRate (ABR) streaming [2]. These improvements go hand in hand with a need for improvements in quality of experience prediction relative to encoding parameters.

This paper aims at finding correlation between predicted quality and user ratings based on three different encoding parameters:

- Resolution ranging from 540p to 4k
- Bitrate ranging from sheisstotal to tomuch
- 7he m4g1c p4r4m3t3r

Several data has been collected in addition to user ratings, these data include:

- Feedback questions
- Behavior during subjective tests
- other?

II. Test Preparation

A. Video Selection

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

1) Preconsiderations: There exists only a few public 4k video databases. They differ in content (scenario, field size, camera movement, illumination condition) as well as in technical aspects (resolution, frame rate, bit rate and color bit depth). To obtain reference video files for the subjective test, parts of the 4k 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 with no cuts inside. 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) Dataset Preparation: For obtaining a good variety of video sequences three data bases are used: The Harmonic [3] which contains 18 different video files, Cable Labs [4] with 9 relevant 4k contents and the Blender Foundation [5] 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 4k 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. Generally, the challenge is to extract 10 seconds sequences from the original video files with no cut inside because there exists only a few.

TABLE I
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

One problem of the video sequences from the database is that even they have a high bit rate and are stored as ProRes, no conclusion referring to the contained visual quality can be done. If the videos were available in RAW, this would be preferred because RAW assures the least compression and no lost of information regarding the meta data. For this reason a preselection with a system, able to play 4k content in ProRes with high bit rates and connected to a 4k screen, is applied. The remaining reference videos are 6 source clips, each with a duration of 10 seconds. All the files have got a resolution of 3840x2160 pixels and are stored in the ProRes format. Because this format is a 10 bit codec only, the original color depth of each video file can not be determined. Further specific informations about the reference videos are shown in Table I, e.g. Air Show, a video from the Harmonic Database, available with 59.94 fps and a bit rate of 1703 Mbit/s. To get out of it a video for the test, a 10 second sequence is extracted that starts at the 48.5 second of the source video file.

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

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

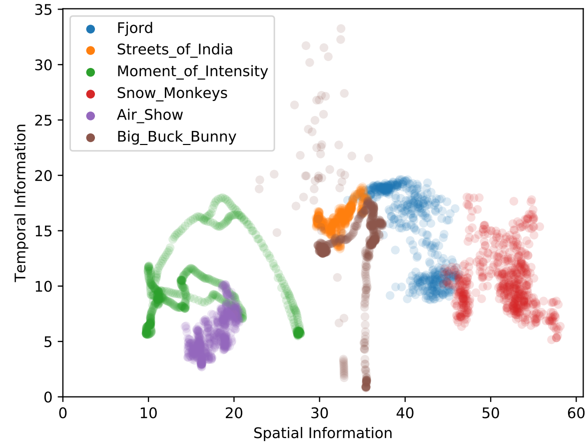

Fig. 2. Spatial and temporal informations of the reference video files. Each cross represents one frame

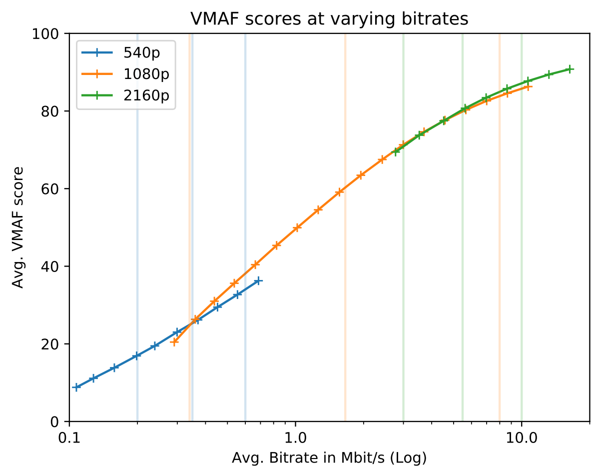

Fig. 3. VMAF scores for 25 different bitrates at 3 resolutions. The encoded bitrates and the VMAF scores are averaged between the 6 sequences and form the abscissa (Log) and ordinate respectively.

Informations about the spatial and temporal changes between the frames of each reference video are calculated with SITI [6], a command-line-based tool that refers to ITU-T P.910, and diagrammed in Figure 2. As expected, Air Show includes the smallest spatial and temporal changes in comparison with the other reference videos due to the slightly changing and low complex scenario. Whereas the cartoon Big Buck Bunny shows a large range of temporal informations by reason of camera movement and Snow Monkeys the highest spatial informations according to fine structures and details.

B. Encoding Parameters

The following section illustrates how the sequences were encoded for the perceptual test and the choice of encoding parameters.

1) Encoding Presets: Two different presets are used for the sequence encodings, a ”naïve” (P.1) and an ”expert” preset (P.2). The ”naïve” preset is a simple CBR encoding, whereas the ”expert” preset is a 2-pass encoding with a Q-CTRL pass followed by a B-CTRL pass. Every sequence is encoded with both presets at 3 resolutions (540p, 1080p, 2160p) and 3 bitrates for each resolution. The resulting VMAF-scores for the encoded sequences can be seen in Figure 4. The following section explains our method for the bitrate selection.

2) Selection of Bitrates: Video Multi-Method Assessment Fusion (VMAF) is a full reference metric for estimating human perception of video quality [7]. We use it because it provides a better estimate of subjective quality than single metrics like SSIM or VIF.

To estimate relevant HEVC 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 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 (foreground) and the target rates (background). The bitrates are chosen at least 2 bitrate-samples away from an intersection with the next quality, except for the lower 1080p-bound where this is not possible due to encoder restrictions, as is fails to encode for even lower bitrates.

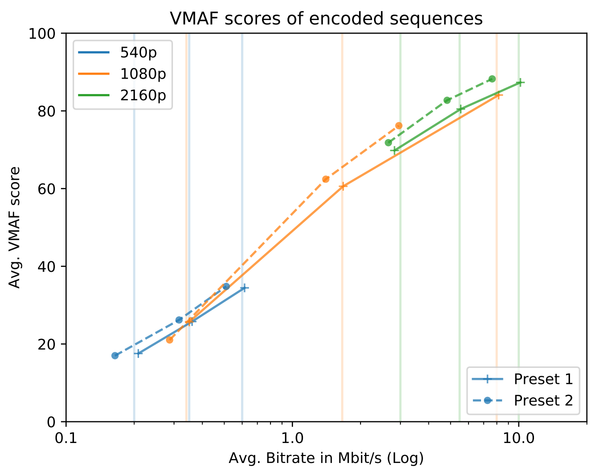

Fig. 4. VMAF scores of encoded videos for both presets. The encoded bitrates and the VMAF scores are averaged between the 6 sequences for each preset and form the abscissa (Log) and ordinate respectively.

3) Encoding Automation: We automate the whole process for downloading, preprocessing and encoding the source videos using pydoit [8]. 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, MPEG-4 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 [7]. 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) [9] and saved as a json file (6). All of the scores are then aggregated into a single CSV for plotting and further analysis (7).

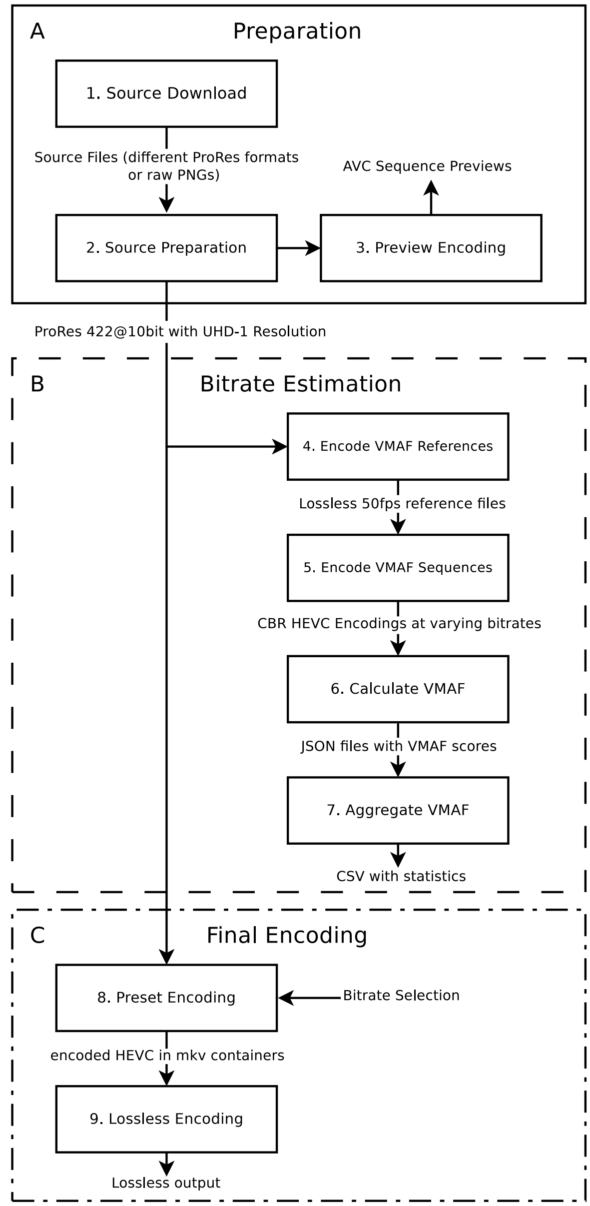

Fig. 5. Automated processing and encoding workflow.

The main encoding takes place last (C). Final encoding requires bitrate selection, which is performed manually. The sequences are encoded with the presets first (8) and then transcoded to a lossless format (ffvhuff) to allow for fast and consistent playback as well as archiving the video material.

C. Test Setup

Several steps have been followed in order to enable the acquisition of meaningful data during the subjective tests. Theses steps included 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.

1) Test Environment: In order to make the results of our research reproducible the ITU P.910 recommendation [10] have been followed. The parameters for the viewing conditions being:

a: Viewing distance
b: Peak luminance of the screen
c: Ratio of luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white
d: Ratio of luminance of background behind picture monitor to peak luminance of picture
e: Chromaticity of background
f: Background room illumination

As performing tests following these specifications is common at the Technical University of Ilmenau, a room meeting these requirements was available and therefore used. The specifications of the room was the following for each of the parameters:

a:
b:
c:
d:
e:
f:

2) Rating Framework: The testing procedure which seems more suited to the case was Absolute Category Rating (ACR) [10], 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 workflow of this rating framework is the following (see fig. 6):

1 Training phase: rating of reference videos
2 Registration phase: submission of personal informations
3 Rating phase: rating of 108 videos
4 End phase: feedback about the test

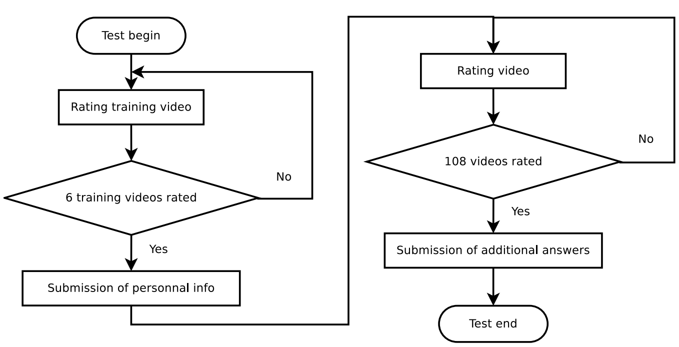

Fig. 6. Rating workflow

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 [10], the following questions have been asked at the end of each rating session:

- Presence of blocky artefacts
- Visible bands of colour
- Smoothness of the playback
- Have you seen 4K content before?
- Have you seen content on a 4k screen before?
- How sure were you about the rating that you provided?

These idea behind these question is dual as these questions may translate how users percieve/are sensitive to video features as well as how users may clearly express their perceptions. As it has already 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 doesn't allow detailed answers some participants may hesitate between two answers and change their answers or hover with their mouse around some answers. Also, answering speed, which could be an indicator of a participant skipping answers or to the contrary being strongly confident of his answers. We believe that several informations can be extracted from these kind of behaviours:

- Confidence in the participant sequence rating
- Confidence in the participant overall rating
- Intermediate scores (eg: 4.5)

III. Test Results

A. Ratings

(MOS Results, Plots and stuff)

B. Features

(Mouse tracking, Additional Questions)

C. Analysis

Similar MOS for naive and expert presets, expert preset is worse for low-bitrate 1080p.

IV. Conclusion

1) Conclude Stuff: 4k Content
MOS is quite alot like VMAF (who would have guessed)
"Expert" Preset is not using available bitrate in critical low-bitrate situations. (Drop-Off for 1080p MOS), can already be predicted from VMAF plot.

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 could provide a better sampling of the bitrate-MOS space and allow for more detailed analysis.

The "expert" preset could be chosen

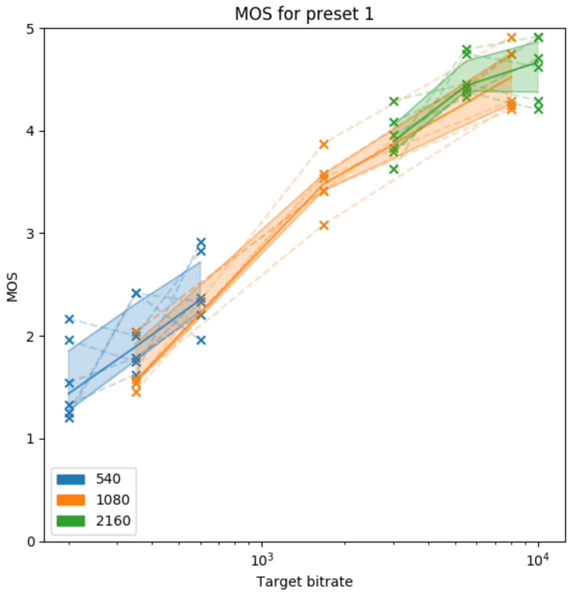

Fig. 7. Correlation between bitrate and MOS for the ”naïve” encoding preset. The center line represents a median and the outer line the 25th and 75th percentile of MOS for the 6 sequences.

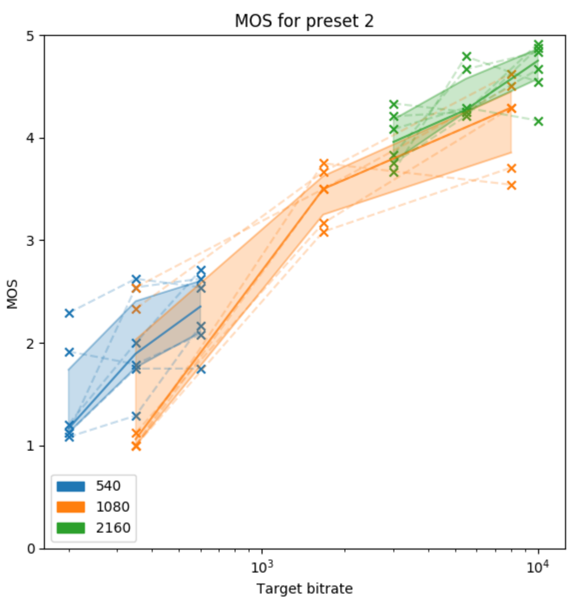

Fig. 8. Correlation between bitrate and MOS for the ”expert” encoding preset. The center line represents a median and the outer line the 25th and 75th percentile of MOS for the 6 sequences.

REFERENCES

[1] D. Farber, R. Greer, A. Swart, and J. Balter, “Internet content delivery network,” Nov. 25 2003, uS Patent 6,654,807. [Online]. Available: https://www.google.com/patents/US6654807

[2] D. Brueck and M. Hurst, “Apparatus, system, and method for multi-bitrate content streaming,” Oct. 19 2010, uS Patent 7,818,444. [Online]. Available: https://www.google.com/patents/US7818444

[3] (2017, Jun) Harmonic demo footage. [Online]. Available: https://www.harmonicinc.com/4k-demo-footage-download/

[4] Cablelabs 4k video. [Online]. Available: http://4k.cablelabs.com/

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

[6] Telecommunication-Telemedia-Assessment. (2016, Dec) Telecommunication-telemedia-assessment/siti. [Online]. Available: https://github.com/Telecommunication-Telemedia-Assessment/SITI

[7] 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.

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

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

[10] I. REC, “P. 910,” *Subjective video quality assessment methods for multimedia applications*, 1998.

Fig. 9. Correlation between mouse related features and rating confidence

Fig. 10. Correlation between feedback related features and rating confidence

Fig. 11. Correlation between bitrate and MOS after correction based on mouse features

Fig. 12. Correlation between bitrate and MOS after correction based on feedback features

Fig. 13. Correlation between bitrate and MOS after correction based on mouse and feedback features