On use of crowdsourcing for H.264/AVC and H.265/HEVC video quality evaluation

Ondrej Zach, Michael Seufert, Matthias Hirth, Martin Slanina, Phuoc Tran-Gia

Angaben zur Veröffentlichung / Publication details:

Zach, Ondrej, Michael Seufert, Matthias Hirth, Martin Slanina, and Phuoc Tran-Gia. 2017. "On use of crowdsourcing for H.264/AVC and H.265/HEVC video quality evaluation." In *27th International Conference Radioelektronika (RADIOELEKTRONIKA), 19-20 April 2017, Brno, Czech Republic*, 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/radioelek.2017.7937581.

On Use of Crowdsourcing for H.264/AVC and H.265/HEVC Video Quality Evaluation

Ondrej Zach*, Michael Seufert†, Matthias Hirth†, Martin Slanina*, Phuoc Tran-Gia†

*Brno University of Technology
SIX Centre, Department of Radio Electronics
Brno, Czech Republic
ondrej.zach@phd.feec.vutbr.cz

†University of Würzburg
Institute of Computer Science
Würzburg, Germany
seufert@informatik.uni-wuerzburg.de

***Abstract*—Crowdsourcing has become a popular method in the field of video quality evaluation. Gathering the opinion of the users using crowdsourcing is quick and relatively cheap but such a study has to be designed very carefully in order to give relevant results. So far, the majority of the QoE studies using crowdsourcing has been focusing on the performance of H.264/AVC algorithm in different situations (such as encoder settings, stalling effects, etc). Modern video coding methods, however, are only rarely tested using the crowdsourcing approach. We designed a study comparing the performance of both H.264/AVC and H.265/HEVC standards in the crowdsourcing environment. We deal with the possibilities of delivering and presenting the HEVC encoded content to the participants of the crowdsourcing study and potential challenges. Finally, the study was performed using Microworkers platform and gathered results are then compared with three different objective video quality metrics.**

I. Introduction

Video streaming and other video-based services represent majority of the overall internet traffic. According to a Cisco study [1], video services take the share of more than 67% of the overall traffic and this share is expected to grow up to 80% in 2019. Furthermore, the users expect to receive the video content at higher visual quality. This has also a significant impact on the amount of the data transferred. To offer higher quality of the video while preserving the same data rate, a change in video compression scheme is necessary.

Currently, the majority of video content on the Internet is encoded using H.264/AVC (Advanced Video Coding). AVC performs best in HD (720p) and Full HD (1080p) scenarios with most often used bit rate values in the range from 2 Mbps to 5 Mbps. For higher resolutions, the compressed bit rate is growing in accordance with the increase of the number of pixels in a frame. For instance, 4K Ultra High Definition (2160p) requires approximately four times the bit rate of full HD as there are four times as many pixels in a frame, which results in a quadruple increase of the encoded bit rate required being in the range of 8 Mbps to 20 Mbps per video stream.

Recently, highly efficient coding techniques have been introduced. One of them is a direct successor to H.264, developed by the Joint Video Team of ISO Motion Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG), standardized jointly as ITU-T H.265 [2] and ISO/IEC 23008-2 [3] and generally denoted as High Efficiency Video Coding (HEVC) standard. This new standard is often stated to need just 50% of the bit rate of AVC while preserving the same visual quality [4].

A majority of codec comparison tests are performed under controlled laboratory conditions, with respect to a prescribed test setup and test procedure, such as [5]. Since the end users of the multimedia services, i.e. the consumers of the video signals, have no controlled conditions available, crowdsourcing tests have the ambition to provide results closer to real setup in general use. Under which circumstances do the viewers really see the difference between AVC and HEVC compressed video sequences? This question is addressed in the following Sections.

II. Related Work

Crowdsourcing is a method, in which the given task is solved by large group of users, usually using an online platform. Using crowdsourcing can easily bring large amount of data/results in short time and therefore it has been widely used in many fields, especially in the field of Quality of Experience, [6], [7].

Traditional testing of perceived image or video quality is performed in a controlled environment according to recommendations of the International Telecommunication Union (ITU) such as Recommendation BT.500 for Standard Definition (SD) Television, [8] or Recommendation P.910 on multimedia applications [5]. Such a study provides relevant results and the impact of the tested settings as perceived by the real viewer. However, gathering enough data takes a lot of time and such a test is often also financially demanding. On the other hand, gathering data using crowdsourcing (CS) is usually both faster and cheaper. Nevertheless, the results from a CS study have to be processed much more carefully - in an uncontrolled environment, it is much easier for the participants to give irrelevant results (accidentally or on purpose), [9].

Studies comparing H.264/AVC and H.265/HEVC using the crowd are very rare. As the purpose of this paper is to design a test and to select appropriate settings of the encoders to be tested, in the following, we describe also subjective studies in controlled environments.

A CS-based study on the performance of AVC, HEVC and VP9 video coding standards is presented by Rerabek, Hanhart et al. in [10]. The authors used the QualityCrowd2 platform in order to create the framework for evaluation of quality of videosequences with HD and FullHD resolution. However, the study was offered in local uncontrolled environment of the

computer lab and the demography of the participants of the test was limited to students and staff of the university. Therefore, the study can not be considered as a real crowdsourcing study.

Objective studies of HEVC are presented by Hanhart et al. and Rerabek et al. in [11], [12], [10]. Their studies mainly compare the performance of HEVC with AVC or VP9 in Ultra HD context. In [13], the authors performed a video quality study on a wide variety of resolutions (from 480p up to UHD) using both H.264 and H.265. A subjective quality study using Degradation Category Rating (DCR) methodology according to [5] was held. The study was performed in controlled environment at two different laboratories and the main focus of the paper is to discuss the quality compared to bit rate savings of HEVC. Studies published in [14], [15] investigate the quality of UHD videos encoded using H.264 and VP9 standards, respectively, compared to HEVC. The quality was measured using objective quality metrics such as peak-signal-to-noise-ratio (PSNR) and Structural Similarity Index (SSIM) only. To sum up, recent studies on video quality of HEVC coded sequences focus on performance of the codec in Full HD and UHD scenarios.

To the best of our knowledge, this is the first study on comparison of the HEVC and AVC codecs focusing on users in their real environment. We present a study with a realistic scenario when the users are watching the videos using their own equipment (PC, laptop). In order to do so, we use crowdsourcing as a tool to acquire a sufficient number of participants. In this scenario we use standard resolution only as streaming higher resolution (1080p, 4K) might not be feasible for many users due to the limitations of either displaying devices or their Internet connections.

III. Study Description

The following paragraphs provide a detailed description of the study setup, content used and preprocessing employed. Also the technical solutions used to overcome the limitations of the web-based crowdsourcing experiment are described.

A. Framework

An online test framework similar to [16] was implemented, such that each participant watched five video sequences with different content each, and rated the visual quality afterwards. To avoid network influences during the playback (e.g., initial delay, stalling), all videos were downloaded to the local browser cache before the playback. The framework followed the best practices described in [9], [17], including monitoring of test execution and reliability checks. To be considered reliable, a user first had to read the test instructions, which also explained a game-like monitor quality pre-test. Also, if the users' display did not meet minimum screen resolution requirements, the user was not able to participate. The clicking behavior during that pre-test already showed if the user read the instructions properly or not. Then, the user had to watch all videos in their full lengths and answer simple content questions correctly. Finally, the test included also personal questions, which were presented twice, at the beginning and at the end of the test, to check the consistency of the answers. If the test was not executed properly (e.g., switching to another browser tab during video playback, wrong answers to simple content questions, different answers to consistency questions), the ratings of the corresponding users were marked as unreliable by the framework and were filtered out later. Furthermore, the framework was able to detect and record several features of the computer system used by the participants (e.g. screen resolution, browser type, OS, download speed etc.), which may be used in our future studies.

In order to focus the participants' attention on the videos, during playback a simplistic web page was shown, which was totally gray and displaying the video player only and the users were not able to resize the player window. For each video content, a condition was selected randomly, but following a water filling algorithm until 11 reliable ratings for each condition were available. After that, an adaptive test design was used to obtain a mean opinion score (MOS) with small confidence intervals, [18] This means, the conditions were selected with a probability proportional to the current size of the 95% confidence interval of the mean of the ratings. Therefore, the number of ratings per clip varied.

The study was available as a micro job on the Microworkers[1] crowdsourcing platform. Every user could participate and was rewarded with 0.30$ upon completion of the test.

B. Browser-based HEVC playback

The main issue we had to face, was how to play the HEVC encoded content on the computers of the crowd workers. A crowdsourcing-based QoE testing should be designed to run smoothly at many different computers with wide variety of configurations. The researcher can rely only on the web browser and its most common features without any additional plugins or extensions. Therefore, any plugin-based solution for HEVC playback had to be omitted. Although Microsoft Edge natively supports HEVC playback, this feature is limited to machines with hardware HEVC decoding only. At the time of writing of this paper, the HW HEVC decoding is available only with the newest generations of CPUs (Intel Skylake or AMD Carrizo, [19], [20]) or most recent GPUs. Relying on this feature would enormously limit the range of the available participants and was not feasible for our case.

To overcome this issue, we decided to re-encode the HEVC encoded sequences once again with AVC. The settings of the AVC encoder were set to minimize the influence of the re-encoding. Using the lossless encoding was not possible again due to the limited support by the web browser. More details on this issue are described in Section III-D.

C. Dataset

For the purposes of subjective quality evaluation a dataset of encoded sequences was created. We used five source sequences, three of them were taken from video databases, the other two were real life sequences. The database sequences (*Basket*, *Bunny* and *Leopard*) were downloaded from the Consumer Digital Video Library[2] and were previously used by National Telecommunications & Information Administration (NTIA) or Video Quality Experts Group (VQEG) in their HD video quality studies. Sequences *Peaky* and *Wacken* were

[1] https://microworkers.com

[2] http://www.cdvl.org

(a) Basketball (b) Big Buck Bunny (c) Leopard (d) Peaky Blinders (e) Wacken

Fig. 1. Source sequences.

acquired from a high-quality blu-ray disc. The sequences are representatives of the majority of the contents that users usually watch online (cartoon, sports, music video), and a more detailed description of the sequences can be found in Tab. I. One frame of each content is shown in Fig. 1.

All source video sequences were available in 1080p resolution at 25 frames per second, the length of the sequences was adjusted to 10 seconds. However, due to the limitations as stated in previous Section, we could use standard resolution (576p) only. Therefore, the source video sequences were downscaled using *ffmpeg*[3] tool to meet this demand. The scaling algorithm used was bicubic for luma component and bilinear for chroma components.

TABLE I. DESCRIPTION OF VIDEO SEQUENCES

NAME	DESCRIPTION	CONTENT CHARACTERISTICS
Basket	Basketball match	Fast motion, camera zoom in/out
Bunny	Big Buck Bunny cartoon	Camera move
Leopard	Leopard in the zoo	No cut, details
Peaky	Peaky Blinders TV series	Slow motion, dark colors
Wacken	Metal concert in Wacken	Fast motion, dark and light, cut

D. Encoding

All sequences were encoded compliant to the H.264 and H.265 video coding standards. For encoding to H.264 and H.265 we used the *x264*[4] and *x265*[5] encoder implementations, respectively. Both encoders were set to medium presets to use their most common features. The x264 encoder was used with setting the size of the transform blocks to 8×8, motion estimation range to 16 pixels. The HEVC encoder was using the settings of CTU size to 64, motion estimation range 57 pixels. Both the x264 and x265 encoders were using deblocking and Sample Adaptive Offset (SAO) filters, respectively.

All source video sequences were encoded to 5 quality levels. These levels were determined by bit rate and the selected values were $\{0.5, 0.8, 1.0, 2.0 \text{ and } 3.0\}$ Mbps. This range of bit rates is commonly used on the Internet for video streaming of Standard Definition video. Together, our database consisted of 50 processed video sequences.

The re-encoding of the HEVC coded processed video sequences (PVS) to AVC was performed in order to provide high playback compatibility on the PCs of the participants with preserving the quality as close to the original PVSs as possible. However, the initial test showed, that the using of lossless encoding to AVC was not feasible due to limited support in web browsers. Therefore, we re-encoded the HEVC PVSs with setting the Quantization Parameter to 1 (the lowest distortion). An expert screening session was conducted to confirm there is no perceivable difference between the original HEVC sequences and the re-encoded sequences as even the expert viewers were not able to tell the difference between the original and re-encoded sequences. Furthermore, the PSNR value computed from original HEVC encoded and AVC re-encoded sequences was higher than 60 dB which further confirms this fact.

However, the drawback of this re-encoding is, that the file sizes of the re-encoded PVSs increased drastically. For example, for the *Basket* content, the file sizes varied from approx. 38 MB to 51 MB for the actual bitrates from 500 kbps to 3 Mbps, respectively.

[3]http://www.ffmpeg.org

[4]http://www.videolan.org/developers/x264.html

[5]http://www.x265.org

IV. RESULTS

The following Section describes both the analysis of gathered data from the point of view of the participant of the crowdsourcing study and the results of video quality metrics. For comparing the quality of the two codecs, we will present both the objective results calculated for the sequences and the results of our subjective study. Objective quality metrics were used in order to verify if there is any measurable difference in the visual quality of compressed sequences.

A. Gathered Data

As each user watched 5 sequences, a high number of participants was envisaged (~1,000 users) in order to get statistically significant results for each encoded sequence. Therefore, the campaign was available online for one month, divided in three runs. As we did not plan to use any crowd limitations (e.g. best earners, geographical limitation) or any user list (e.g. a list of reliable users based on previous studies), each user (identified by his unique Microworkers id) could theoretically participate in the study three times. Altogether, the test was completed 903 times by 486 unique users from 65 different countries, most of them originating from Bangladesh (97), Belgium (76), Serbia (44), and Romania (42). Following the strict consistency checks of the framework, we observe a consistency rate of 47.29%, which means that the participants did not conduct the study properly in slightly more than half of the tests. The scores gathered from these users were not used in the further processing of the data. Our monitoring also indicated that some users experienced stops during playback of the sequences (for both H.264 and H.265 encoded sequences), which were probably caused by playback problems in the browser. We conducted a t-test, which showed that there is a significant difference between the ratings for disturbed sequences (sequences with stops during playout) and undisturbed sequences (p value in the order of 10^{-12}). Therefore, these ratings were

also omitted from the final evaluation of the data. After filtering out unreliable users and ratings from disturbed sequences, we had 1398 scores altogether, which correspond to 27 scores per encoded sequence on average.

For possible future studies, we also monitored the type of the web browser the participant use. More than 62% of the participants used Google Chrome. On the 2nd place was Mozilla Firefox with 28%. Only in 4 cases, we detected the use of Microsoft Internet Explorer, which includes also previously mentioned Microsoft Edge with built-in HEVC decoding support.

B. Objective Quality Assessment

In order to create a repeatable and objective comparison of the encoded video sequences used throughout the subjective crowdsourcing experiment, we evaluate the most commonly used full reference objective video metrics. These metrics capture the differences in pixel sample values (PSNR) or frame image structures (SSIM). Further, one metric based on a human visual system model (VQM) is supposed to create the most relevant reference for repeatable video quality comparison.

The PSNR (Peak Signal-to-Noise Ratio) is plotted for all sequences in Fig. 2 grouped by content and ordered by bit rate. It can be seen that the different contents result in different PSNR ranges for the compressed sequences, which again indicates the diverse characteristics of the videos. For each content, as expected, the PSNR increases with increasing bit rate. However, when directly comparing H.264 clips (light blue bars) to H.265 sequences (dark blue bars), H.265 is able to always achieve a higher PSNR with the same bit rate. This effect is the strongest for Leopard and Wacken sequences, which have generally lower PSNR level for the H.264 sequences as they required highly lossy compression to reach the target bit rates.

In contrast to PSNR, SSIM (Structural Similarity, [21]) was designed to measure the similarity between pictures taking into account phenomena of human visual perception. Fig. 3 plots the SSIM index, which has a codomain from 0 to 1, for all tested sequences. Again, it can be seen that encoding with H.265 always improves the visual quality compared to the corresponding H.264 sequence with the same bit rate and has a higher SSIM value. In particular, the quality of low bit rate videos can be substantially improved by using H.265. For example, for the Wacken sequence at 500 kbps, the H.265 clip has a higher visual quality than the H.264 clip at 800 kbps, and even reaches SSIM index value similar to the H.264 clip at 1000 kbps. This means, in this case, H.265 achieves the same visual quality with half the bit rate required by H.264. This is in accordance with [22], where the authors observed similar bit rate savings.

Finally, we also compare the performance of H.264 and H.265 with VQM (video quality metric, [23]), which is a standardized method to compute the video quality using perception-based features ranging from 0 to 1. The objectively measured quality of all sequences can be seen in Fig. 4, which has a reversed y-axis to reflect the fact that lower VQM values indicate a higher quality. The plots indicate that the same qualitative observations also hold for VQM. Comparing H.264 and H.265 sequences of the same bit rate, H.265 is

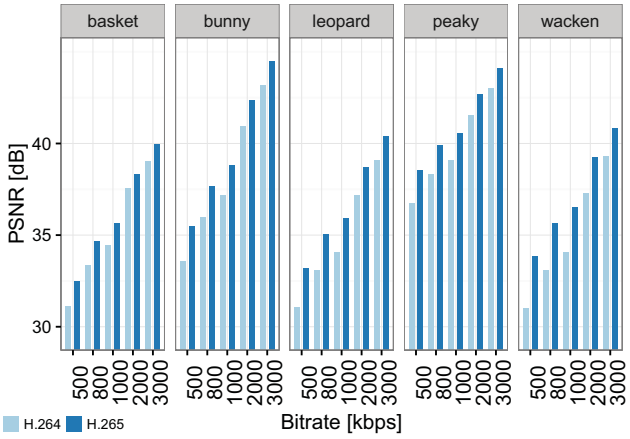

Fig. 2. Objectively measured quality using PSNR.

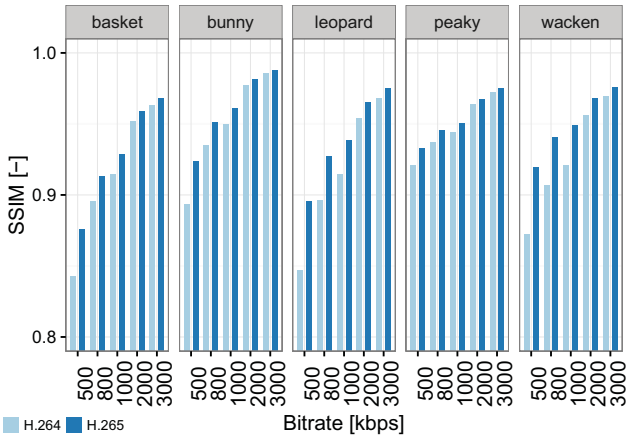

Fig. 3. Objectively measured quality using SSIM.

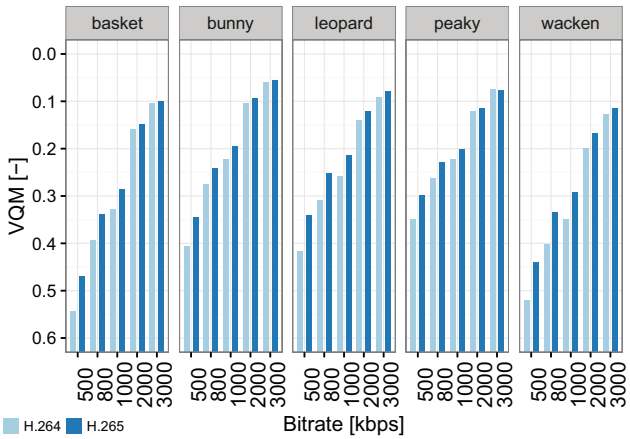

Fig. 4. Objectively measured quality using VQM.

able to reach a lower VQM value, and thus, higher visual quality. Nevertheless, in terms of the VQM metric, the quality gains are not as large as measured by SSIM. For example, when revisiting the same Wacken sequence at 500 kbps already discussed for the SSIM metric, VQM indicates that the H.265 clip does not reach a higher visual quality than the H.264 clip at 800 kbps. This contradicts the judgment based on SSIM that

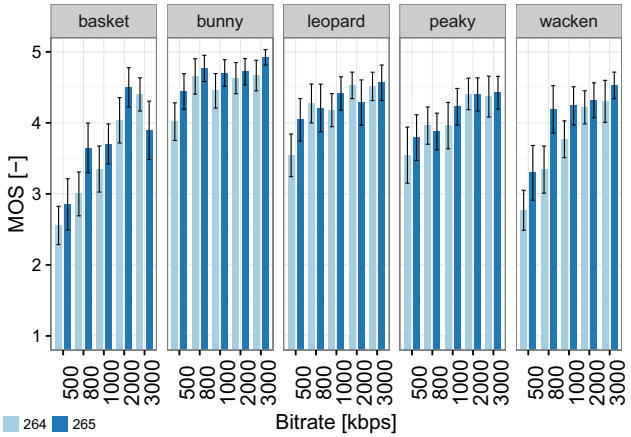

Fig. 5. MOS and 95% confidence intervals of subjective assessments.

H.265 at 500 kbps has a higher visual quality than H.264 at 800 kbps.

To sum up, although the objective metrics already help to infer that encoding with H.265 improves the visual quality compared to H.264 for a given target bit rate, the results do not allow for a quantitative comparison of the two codecs. The different objective metrics were not capable to unanimously measure the gain of H.265 over H.264. Thus, in the following, the two encoding methods will be investigated based on subjective quality assessment.

C. Subjective Quality Assessment

Figure 5 shows the results of the subjective crowdsourcing study. The horizontal axis depicts the sequences grouped by content and ordered by bit rate. Light blue bars represent H.264 clips and dark blue bars indicate H.265 clips. The MOS, i.e., the average rating of all participants, is plotted on the vertical axis including the 95% confidence intervals. Before investigating the subjective results in detail, we compare the results of objectively and subjectively measured quality. Therefore, Table II lists the correlation of the objective metrics and the MOS in terms of Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SROCC). In general high correlations are visible with absolute values of at least 0.7. The highest Pearson correlation coefficients are obtained by the SSIM metric and the highest Spearman correlation coefficients can be reached with the PSNR metric. However, differences in the correlation between the used objective metric are rather small compared to the differences between contents. This means that particular characteristics of the content cannot be captured well by the objective metrics.

Taking a look at the MOS ratings in Figure 5, we see an increase of the MOS for each content and codec, when the bit rate increases. The only exception can be seen with the basket sequence encoded using H.265 for bit rates 2Mbps and 3Mbps, however, the confidence intervals of the corresponding MOS values overlap. The influence of bit rate is significant with p value of 0.0101 for a t-test on all ratings. However, the Big Buck Bunny clip, which is an animated cartoon clip, reaches high ratings almost regardless of the bit rate. This is due to the fact that usually higher compression ratios can be achieved when processing cartoon-like content with preserving good visual quality. A similar observation holds for the Peaky Blinders clip and its rather dark images. In contrast, a clear differentiation of the bit rates can be observed for the Basket and the Wacken clip, which feature fast motion and sharp contrasts, respectively. The different perception of different contents can be validated by a t-test on the subjective ratings, which gives a p value of 0.0002. This different perception cannot be accurately revealed by the objective metrics, which results in the lower correlations reported in Table II. This confirms our assumption that objective metrics cannot well capture the complex human perception, thus, the use of current methods for objective QoE estimation is limited.

TABLE II. CORRELATION OF OBJECTIVE METRICS AND MEAN OPINION SCORES.

CONTENT	PLCC			SROCC		
	PSNR	SSIM	VQM	PSNR	SSIM	VQM
Basket	0.92	0.94	-0.94	0.92	0.90	-0.90
Bunny	0.77	0.85	-0.78	0.78	0.78	-0.70
Leopard	0.83	0.91	-0.88	0.87	0.90	-0.90
Peaky	0.95	0.97	-0.97	0.90	0.93	-0.94
Wacken	0.90	0.95	-0.90	0.85	0.85	-0.87

When looking at the difference between H.264 and H.265 codecs, also a less clear result can be inferred. Although most H.265 clips reach a higher MOS than their H.264 counterpart (three sequences show reversed results), the confidence intervals nearly always overlap. This means that we cannot immediately see if the results for the two codecs are significantly different. Therefore, we again conduct a t-test to compare the two codecs. The p value is 0.0782, and thus, cannot be considered significant for a typical confidence level of 5%. This means, that H.265 achieves only a slight visual improvement over H.264, which is not valued by the end users. This is in accordance with results in [24], where the authors were not able to tell clear benefits of HEVC for mobile environments with lower resolutions. As our study settings are already covered by H.264, the quality improvements introduced by H.265 are therefore smaller as for other scenarios. Additionally, the small SD resolution of the test videos could make it hard for users to see differences in the visual quality. Thus, in future work, the study has to be repeated for different bit rates and resolutions.

V. CONCLUSION

In this paper, we designed an experiment to examine the possibilities of using crowdsourcing to evaluate the perceived video quality when streaming H.264 and H.265 encoded videos. Crowdsourcing approach allowed for gathering the opinions of real viewers, as the participants watched the evaluated videos in their natural environment using the equipment (PC, laptop) they are used to. As the traditional subjective quality assessments are commonly held in controlled laboratory environment, such an approach can bring results closer to real situation.

The CS study was then offered to the participants on the Microworkers platform. However, during both the design and the run of the study, several drawbacks of using crowdsourcing to compare the impact of AVC and HEVC encoding algorithms have arisen. The main limitation was the playback of the HEVC encoded content on the devices of the participants. The

direct playback was not possible as such a solution currently relies either on browser plugins or direct HW encoding support by the system. Neither of these solutions is feasible in crowdsourcing environment. The re-encoding solution we proposed has on the other hand limitations in much higher amount of data to be downloaded, which can also influence the number of the available participants of the study. Furthermore, this also limited the study to SD resolution only, as the file sizes of re-encoded files with higher resolutions would be unacceptable.

Next goal of the study was to investigate, if the users can perceive a difference in quality when watching videos encoded using the most recent video coding algorithm H.265/HEVC. Video database used in this study consisted of 5 different video contents to cover the majority of the videos streamed online with 5 different bit rate levels. To offer a complex evaluation, objective quality metrics were used along the subjective assessment.

In general, in this simulated typical usecase, H.265 achieved only a slight visual improvement over H.264, which is not valued by the end users. However, as the study was limited to SD resolution only due to the use of crowdsourcing environment, the results may differ for higher resolutions. On the other hand, even for SD resolution a difference of subjectively measured quality can be seen, therefore using H.265 can bring bit rate savings with preserving similar quality compared to H.264. A remaining issue hindering immediate practical application is the lack of support of H.265 by current browsers, which should evolve soon to benefit from the streaming of H.265 videos.

ACKNOWLEDGMENT

The research published in this paper was financially supported by the BUT Internal Grant Agency under project no. FEKT-S-17-4426 and by the National Sustainability Program under grant LO1401. This work was also supported by the Deutsche Forschungsgemeinschaft (DFG) under grants HO4770/1-2 and TR257/31-2 (OekoNet) as well as grants HO4770/2-1 and TR257/38-1 (Crowdsourcing). The authors alone are responsible for the content.

REFERENCES

[1] Cisco, "The Zettabyte Era: Trends and Analysis," Cisco, Tech. Rep., 2015.

[2] International Telecommunication Union, "ITU-T Recommendation H.265: High efficiency video coding," 2015.

[3] International Standards Organization/International Electrotechnical Commission (ISO/IEC), "23008-2:2015 Information Technology – High Efficiency Coding and Media Delivery in Heterogeneous Environments – Part 2: High Efficiency Video Coding," 2015.

[4] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec 2012.

[5] International Telecommunication Union, "ITU-T Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications," 2008.

[6] E. Estellés-Arolas and F. González-Ladrón-De-Guevara, "Towards an integrated crowdsourcing definition," *Journal of Information science*, vol. 38, no. 2, pp. 189–200, 2012.

[7] M. Hosseini, K. Phalp, J. Taylor, and R. Ali, "The four pillars of crowdsourcing: A reference model," in *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, May 2014, pp. 1–12.

[8] International Telecommunication Union, "ITU-T Recommendation BT.500: Methodology for the subjective assessment of the quality of television pictures," 2012.

[9] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[10] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Quality Evaluation of HEVC and VP9 Video Compression in Real-time Applications," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, May 2015, pp. 1–6.

[11] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective Quality Evaluation of the Upcoming HEVC Video Compression Standard," *Proc. SPIE*, vol. 8499, pp. 84 990V–84 990V–13, 2012.

[12] M. Rerabek and T. Ebrahimi, "Comparison of Compression Efficiency between HEVC/H.265 and VP9 Based on Subjective Assessments," *Proc. SPIE*, vol. 9217, pp. 92 170U–92 170U–13, 2014.

[13] T. Tan, R. Weerakkody, M. Mrak, N. Ramzan, V. Baroncini, J. Ohm, and G. Sullivan, "Video Quality Evaluation Methodology and Verification Testing of HEVC Compression Performance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. 99, pp. 1–1, 2015.

[14] M. Uhrina, L. Sevcik, J. Frnda, and M. Vaculik, "Impact of H.264/AVC and H.265/HEVC Compression Standards on the Video Quality for 4K Resolution," *Advances in Electrical and Electronic Engineering*, vol. 12, no. 4, pp. 368–376, 2014.

[15] ——, "Impact of H.265 and VP9 compression standards on the video quality for 4K resolution," in *Telecommunications Forum Telfor (TELFOR), 2014 22nd*, Nov 2014, pp. 905–908.

[16] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of YouTube QoE via Crowdsourcing," in *Proceedings of the IEEE International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE)*, Dana Point, CA, USA, 2011.

[17] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force Crowdsourcing," 2014, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).

[18] M. Seufert, O. Zach, T. Hofeld, M. Slanina, and P. Tran-Gia, "Impact of test condition selection in adaptive crowdsourcing studies on subjective quality," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, June 2016, pp. 1–6.

[19] Intel, "Learn about the Significance of HEVC (H.265) Codec," Intel, Tech. Rep., 2015. [Online]. Available: https://software.intel.com/en-us/blogs/2015/12/11/codecs-are-they-slowing-you-down

[20] AMD, "AMD Unveils 6th Generation A-Series Processor," AMD, Tech. Rep., 2015. [Online]. Available: http://www.amd.com/en-us/press-releases/Pages/amd-unveils-6th-eneration-2015jun02.aspx

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, "Performance Comparison of H. 265/MPEG-HEVC, VP9, and H. 264/MPEG-AVC Encoders," in *Proc. of the 30th Picture Coding Symposium (PCS)*, San Jose, CA, USA, 2013.

[23] M. H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[24] R. Garcia and H. Kalva, "Subjective evaluation of hevc and avc/h.264 in mobile environments," *IEEE Trans. Consum. Electron.*, vol. 60, no. 1, pp. 116–123, February 2014.