

Comparison of compression efficiency between HEVC/H.265 and VP9 based on subjective assessments

Martin Řeřábek and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG),
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

ABSTRACT

Current increasing effort of broadcast providers to transmit UHD (Ultra High Definition) content is likely to increase demand for ultra high definition televisions (UHDTVs). To compress UHDTV content, several alternative encoding mechanisms exist. In addition to internationally recognized standards, open access proprietary options, such as VP9 video encoding scheme, have recently appeared and are gaining popularity. One of the main goals of these encoders is to efficiently compress video sequences beyond HDTV resolution for various scenarios, such as broadcasting or internet streaming. In this paper, a broadcast scenario rate-distortion performance analysis and mutual comparison of one of the latest video coding standards H.265/HEVC with recently released proprietary video coding scheme VP9 is presented. Also, currently one of the most popular and widely spread encoder H.264/AVC has been included into the evaluation to serve as a comparison baseline. The comparison is performed by means of subjective evaluations showing actual differences between encoding algorithms in terms of perceived quality. The results indicate a dominance of HEVC based encoding algorithm in comparison to other alternatives if a wide range of bit-rates from very low to high bit-rates corresponding to low quality up to transparent quality when compared to original and uncompressed video is considered. In addition, VP9 shows competitive results for synthetic content and bit-rates that correspond to operating points for transparent or close to transparent quality video.

Keywords: High Efficiency Video Coding (HEVC), Advanced Video Coding (AVC), VP9, WebM, subjective quality assessment, coding efficiency

1. INTRODUCTION

Nowadays, there is an increase of video consumption in UHD format, which is distributed through the broadcast channels, digital networks, or packaged media. Such an emerging trend creates a bigger demand on the development of video acquisition, compression, and display technologies, as well as of network and transmission media capabilities. One of the most critical requirement arises in video compression domain, where it is necessary to satisfy the need for a highly efficient video coding standard. ITU-T and ISO/IEC jointly developed video compression standard H.265/MPEG-H Part 2,¹ referred to as HEVC (High Efficiency Video Coding) in the rest of the paper, has been released recently. HEVC was designed to replace its predecessor H.264/MPEG-4 Part 10,² referred to as AVC (Advanced Video Coding) in the rest of the paper, in almost all applications with specific focus on encoding of UHD video. Both standards were jointly developed by group of video coding experts from ITU-T Study Group 16 (VCEG) and ISO/IEC JTC 1/SC 29/WG 11 (MPEG), called the Joint Collaborative Team on Video Coding (JCT-VC)*.

As mentioned above, HEVC was designed to target a broad scale of applications such as mobile TV, home cinema, and especially ultra high definition television. HEVC supports state-of-the art and next-generation of display technologies offering higher resolutions, frame rates, and improved picture quality in terms of noise level, color gamut, and dynamic range. Most recent video coding related achievements in research and industrial field

Further author information: (Send correspondence to Martin Řeřábek)

Martin Řeřábek: martin.rerabek@epfl.ch

Touradj Ebrahimi: Touradj.ebrahimi@epfl.ch

*<http://www.itu.int/ITU-T/studygroups/com16/jct-vc/>

have been implemented with HEVC as it has been reported that it achieves a dramatic bit-rate reduction in comparison to AVC.

In addition of international video coding standardization processes of ITU-T and ISO/IEC, Google[®] proposed an open, royalty-free, video file format WebM* specifically designed for web applications. Video data within WebM format are encoded by proprietary video codecs VP8 or VP9. VP8 codec³ was originally developed by On2 Technologies[®] Inc., which was acquired by Google[®] Inc. in early 2010. Further development of codec for WebM file format resulted in next-generation encoding scheme VP9 as an official successor of VP8 codec. The explicit goal of VP9, recently announced to be finalized, was to produce much more compact bitstreams than VP8 especially for high definition content, with only a modest increase in decoding complexity.⁴

Whereas comparison between HEVC and AVC in terms of coding efficiency has already been investigated^{5,6} extensively based on both, objective measures and perceived quality, less is known about coding efficiency of VP9 especially for UHD TV content. It was shown, that for resolution beyond HDTV, HEVC achieves substantially better performance in comparison to AVC and it allows to reduce bit-rate by up to 65% while maintaining the same perceived quality.⁶ Furthermore, in a recent study of Grois et al.,⁷ evaluation of VP9 in terms of coding efficiency for up to WQHD resolution content is presented and an average bit-rate overhead of 8.4% and 79.4% is declared in comparison to AVC and HEVC based on objective metrics. However, to the authors best knowledge, no comparison of the above mentioned encoding algorithms in terms of perceived quality based on subjective evaluation has been performed and presented in the literature.

In this paper, a performance benchmark of AVC, HEVC and VP9 encoders for UHD TV resolution is presented, assuming broadcast application scenario. Comparison of encoders is based on results and detailed analysis of a subjective quality evaluation campaign performed on a professional high-performance 4K/Quad Full High Definition (QFHD) LCD reference monitor according to ITU recommendation.

The paper is organized as follows. In the next section, the selected coding algorithms compared in our experiments and their main configuration parameters are introduced and briefly described. Section 3 contains detailed description of the selected test video sequences, the testing environment, the test methodology adopted in our experiment, and brief description of collected data preprocessing. Statistical analysis of the collected subjective data and the results are presented in Section 4. Finally, concluding remarks are drawn in Section 5.

2. CONFIGURATION OF SELECTED CODECS

For AVC based encoding,^{2,8} the JM reference software encoder[†] was selected and it serves as an anchor for comparison of next-generation codecs HEVC and VP9. On the other hand, HM reference software[‡] encoder was selected as the most popular encoding implementation for HEVC based encoding.⁹ As a third encoding scheme, the latest result of the WebM Project, i.e. VP9 codec[§] released by Google[®] Inc. was selected and used in our experiments. This section briefly describes the configuration of selected encoders and their profiles.

2.1 AVC and HEVC profiles

The latest versions of JM and HM reference software models (JM 18.6 and HM 15.0) were used for encoding video sequences with AVC and HEVC. The Random Access (RA) configuration was selected for both JM and HM reference encoders since it gives better results than the Low Delay (LD) configuration. The Group of Pictures (GOP) size was set to 8 pictures and the Intra Period was set to 24 and 32 pictures for 24 and 30 fps contents, respectively. Hierarchical B-pictures were used, with 12% quantization step size increase (corresponds to quantization parameter (QP) increase of 1) between each Temporal Level. The Coding Order was set to 0 8 4 2 1 3 6 5 7. The configuration parameters for AVC and HEVC were selected according to.⁶ More details on the configurations can be found in Table 1.

*<http://www.webmproject.org/>

†<http://iphone.hhi.de/suehring/tml/>

‡https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

§<http://www.webmproject.org/vp9/>

Codec	AVC	HEVC
Encoder	JM 18.6	HM 15.0
Profile	High	Main
Level	5.1	6.2
Reference Frames	4	4
R/D Optimization	On	On
Motion Estimation	EPZS	TZ
Search Range	128	64
Group of Pictures	8	8
Hierarchical Encoding	On	On
Temporal Levels	4	4
Intra Period	1s	1s
Deblocking Filter	On	On
8x8 Transform	On	-
Coding Unit size / depth	-	64 / 4
Transform Unit size min / max	-	4 / 32
Rate Control	Off	Off
Internal Bit Depth	8	8

Table 1: Selected settings for the JM and HM reference software encoders.

2.2 VP9 encoder

The VP9 video codec includes similar coding tools as in AVC and HEVC, and it offers some alternative tools such as adaptive mixing strategies for artificial reference frames, processor adaptive real time encoding, or a low complexity loop filter. For this work, the implementation of WebM Project VP9 encoder and decoder version 1.2.0 – 5183 – *ge002bb9* was considered. Unfortunately, there is no official specification of VP9 encoder nor a manual for this encoder to define the encoding and decoding parameters. Therefore, the setting of VP9 encoder and decoder parameters was selected and decided based on recommendation of the WebM Project leading developers. The Intra Period parameter was set to one second as well, which corresponds to AVC and HEVC based codec settings. Since JM and HM reference encoder, unlike VP9 encoder, don't dispose an adaptive quality mode, that can improve visual quality in some cases, this feature was turned off in VP9 encoder (i.e. `-aq-mode` parameter was set to zero). The final configuration of VP9 encoding parameters allowing the comparative testing versus AVC and HEVC encoders was selected and it is presented in Table 2. To obtain more detailed information about VP9 commands, the reader should refer to an encoder parameter guide available online[¶].

For all three encoders, the Intra Period parameter was set in order to simulate a broadcast scenarios, where a high frequency of I frame patterns is required from several reasons, such as, for example, a channel switching. To compare all coding algorithms in a fair manner, the fixed QP should be set in each of them. Setting the fixed quality parameter for each codec separately allows their fair mutual comparison since it removes all rate control adaptation between video frames. Whereas, there is a simple unique procedure to change the coding quality factor for both JM and HM reference encoders, VP9 encoder allows to fix the QP in two different ways. First approach to fix quality factor in VP9 encoder, used by Grois et al.,⁷ sets the `--min-q` and `--max-q` parameter to the same value of QP. However, according to the comments of lead developers of VP9, such a setting apparently disables all advantages of forward alternative reference frames in VP9 encoder and breaks usability of multiple prior reference frames. It also forces all blocks to have the same rate-distortion (RD) model so the compression efficiency could suffer. Therefore, to achieve an alternative comparison to what has already been reported in literature, the available fixed quality mode `--end-usage=3`, which allows to vary the coding quality factor, was selected for VP9 encoding.

[¶]<http://www.webmproject.org/docs/encoder-parameters/#2-pass-best-quality-vbr-encoding>

```

--good --cpu-used=0
--codec=vp9 -v -t 0 -w < Width > -h < Height > -p 2
--aq-mode=0
--fps=< FrameRate >
--end-usage=3 --cq-level=< QP >
--kf-max-dist=< IntraPeriod > --kf-min-dist=< IntraPeriod >
--minsection-pct=0 --maxsection-pct=2000
--auto-alt-ref=1
--arnr-maxframes=7 --arnr-strength=5

```

Table 2: Selected setting for VP9 encoders.

3. SUBJECTIVE EVALUATION

3.1 Dataset preparation and description

The number and availability of publicly accessible and agreed datasets in UHD for research purposes is still quite limited. There are only few datasets of UHD high quality uncompressed content available. For our tests we selected three different natural contents. Two video sequences available for HEVC verification tests, which took place in early 2014, namely *Manege*^{||} and *Traffic*^{**} were selected. Further, *Tree shade* sequence from SJTU dataset¹⁰ was selected. To cover a wider application scenario we also selected synthetic content from Sintel, which is a computer generated movie produced by the Blender Institute using the open source computer graphics software Blender and released under the Creative Commons Attribution license^{††}. Two synthetic scenes were included in the dataset, one for the test *Sintel2* and one for the training *Sintel39*. The dataset is thus composed of five contents, one for the training and four for the actual tests, with different visual characteristics, resolutions, and frame rates, as follows:

1. *Manege*, 3840×2160 , 30fps
2. *Sintel2*, 3840×1744 , 24fps
3. *Traffic*, 3160×2048 , 30fps
4. *Tree shade*, 3840×2160 , 30 fps
5. *Sintel39*, 3840×1744 , 24fps, used for training

It should be noted that the original frame rate of content *Manege* is 60fps. Nevertheless, due to the rendering limitations explained further, it was converted to 30fps by selecting each odd frame only. The representative sample frame of each content is shown in Figure 1.

Analysis of spatial (SI) and temporal (TI) perceptual information¹¹ was performed and Figure 2 shows the SI and TI indexes on the luminance component of each content. It can be observed that the SI and TI indexes vary from relatively small to relatively large values for the selected content. All test sequences were stored as raw video files, progressively scanned, with YUV 4:2:0 color sampling, and and represented at 8 bits per sample.

For each content and codec, five different bit-rates were selected corresponding to various video quality. Targeted bit-rates were selected separately for each content due to the different spatial and temporal perceptual information and the presence of both natural and synthetic content. Assuming the broadcast scenario usage, it was decided to set the upper limit of the targeted bit-rate to 20 Mbps. As mentioned above, fixed QPs were selected and used to control quality level of AVC, HEVC, and VP9 encoded bitstreams. Whereas the typical

^{||}4EVER consortium[©] - <http://4ever-project.com/>

^{**}Plannet Inc.[©]

^{††}<http://www.sintel.org>



Figure 1: Sample frames of individual contents used as testing (a)-(d) and training (e) stimuli in the subjective evaluation campaign.

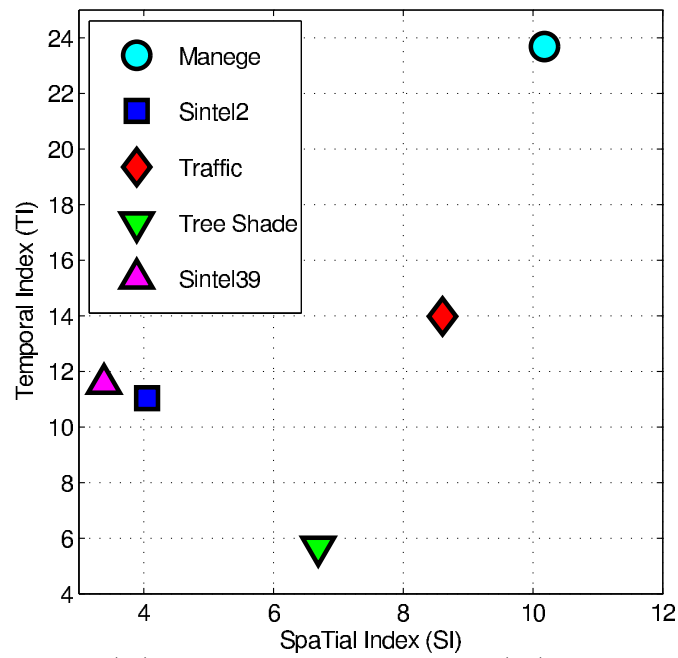


Figure 2: Spatial information (SI) versus temporal information (TI) indexes of the selected contents.

Content	Codec	R1'	R1	R2'	R2	R3'	R3	R4'	R4	R5'	R5
Manege	AVC	5.000	6.877	7.000	8.216	10.000	9.495	15.000	15.235	20.000	19.671
	HEVC	5.000	5.199	7.000	6.758	10.000	9.834	15.000	14.760	20.000	19.248
	VP9	5.000	4.825	7.000	7.062	10.000	10.375	15.000	15.505	20.000	20.239
Traffic	AVC	3.500	3.470	5.000	4.904	7.000	7.202	10.000	9.424	15.000	14.709
	HEVC	2.500	2.302	3.500	3.399	5.000	5.087	7.000	6.862	10.000	10.752
	VP9	3.500	3.693	5.000	5.488	7.000	6.847	10.000	9.803	15.000	14.571
Tree Shade	AVC	5.000	4.942	7.000	7.774	10.000	10.450	15.000	14.192	20.000	19.542
	HEVC	5.000	3.954	7.000	7.245	10.000	9.700	15.000	15.496	20.000	21.684
	VP9	5.000	5.276	7.000	7.224	10.000	9.958	15.000	15.124	20.000	20.521
Sintel2	AVC	1.200	1.289	2.000	1.1777	2.500	2.243	3.500	3.310	4.000	4.375
	HEVC	0.500	0.453	1.000	0.995	1.500	1.542	2.500	2.553	3.500	3.540
	VP9	0.500	0.565	1.000	1.086	1.500	1.623	2.500	2.501	3.500	3.694

Table 3: Targeted R' and actual R bit-rates (Mbps).

Content	Codec	R1	R2	R3	R4	R5
Manege	AVC	38	36	35	31	29
	HEVC	38	36	33	30	28
	VP9	63	61	57	51	44
Traffic	AVC	37	34	31	29	26
	HEVC	38	35	32	30	27
	VP9	63	61	59	55	44
Tree Shade	AVC	33	30	28	26	24
	HEVC	34	30	28	25	23
	VP9	60	57	53	41	36
Sintel2	AVC	32	29	27	24	22
	HEVC	34	28	25	22	20
	VP9	52	40	34	28	23

Table 4: Quantization Parameters (QPs) for all encoding schemes.

QPs for AVC and HEVC are in the range of 25 to 37 the QP values, for VP9 encoders they may vary from 0 to 63. First, several sequences were encoded for each content and codec using QP values within the typical ranges given above, and attempting to correspond the QPs from one codec to another. Afterwards, an expert screening session was conducted to select the lower and upper QP bounds for each content separately, targeting realistic bit-rates to cover the full quality scale. Finally, the targeted bit-rates were refined and validated during a second expert screening session. The training material was selected during the last expert screening session to cover the full quality scale. For the three intermediate quality levels, examples of both AVC, HEVC and VP9 degradations with similar strengths were selected. The complete sets of targeted ($R1'$ - $R5'$) and actual ($R1$ - $R5$) bit-rates are shown in Table 3. The final QP values used to encode selected testing video sequences are specified in Table 4.

3.2 Testing environment and equipment

To avoid the involuntary influence of external factors and to ensure the reproducibility of results, the laboratory for subjective video quality assessment was set up according to ITU recommendation BT.500-13.¹² Professional high-performance 4K/QFHD LCD reference 56-inch monitor Sony Trimaster SRM-L560^{††} was used to display

^{††}http://pro.sony.com/bbsccms/assets/files/cat/mondisp/brochures/di0195_srm1560.pdf



Figure 3: Testing room setup.

video stimuli. The monitor consists of four Full HD panels, which are driven by four display ports and mutually synchronized by the graphic board of the video server to prevent any tearing effect. To drive the monitor, the professional graphic board ATI FirePro V8800 is used. This configuration allows to operate the monitor in QFHD mode providing QFHD native resolution. The monitor was calibrated using an EyeOne Display2 color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level. The room was equipped with a controlled lighting system that consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid grey. The illumination level measured on the screens was 15 lux and the ambient black level was 0.2 cd/m^2 . The distance of the subjects from the monitor was approximately equal to 1.6 times the height of the displayed content as required in ITU Recommendation ITU-R BT.2022,¹³ which in our case corresponds to 1.12m. A picture of the MMSPG test laboratory serving as a evaluation environment is shown in Figure 3. Rendering the raw 2160p 30 fps YUV 4:2:0 video sequences in native spatial and temporal resolutions, requires data-rate of 373.25 MB/s. Therefore, a special hardware exploiting the capability of Solid State Drives (SSD) was adopted.

3.3 Test methodology

In a subjective quality assessment test, a set of video sequences is presented in a predefined order to a group of subjects, who are asked to rate their visual quality on a particular rating scale. To compare the quality of different encoding schemes, the Double Stimulus Impairment Scale (DSIS) method,¹² Variant II, with a continuous impairment scale was chosen. This method is useful for assessing the quality of test material with impairments which are difficult to detect. In a DSIS test, pairs of sequences, i.e., stimuli A and B, are sequentially presented twice to the subject and she/he is asked to rate the quality of the second stimulus, as shown in Figure 4(a). The subject is told about the presence of the reference video (stimulus A), having the best expected quality, and she/he is asked to rate the level of annoyance of the visual defects that is observed evaluated video stimuli (stimulus B). The annoyance is divided into five different levels labelled as *Very annoying*, *Annoying*, *Slightly annoying*, *Perceptible*, and *Imperceptible*. The used rating scale is shown in Figure 4(b).

The evaluation task requires a lot of attention and concentration, therefore the subjective test should be split into separate sessions no longer than 30 minutes each. Each Session should be followed by a resting phase where subjects relax their visual system. Furthermore, in order to retain the concentration of the subjects, it is recommended that as many different contents as possible is alternated within the same session. To prevent

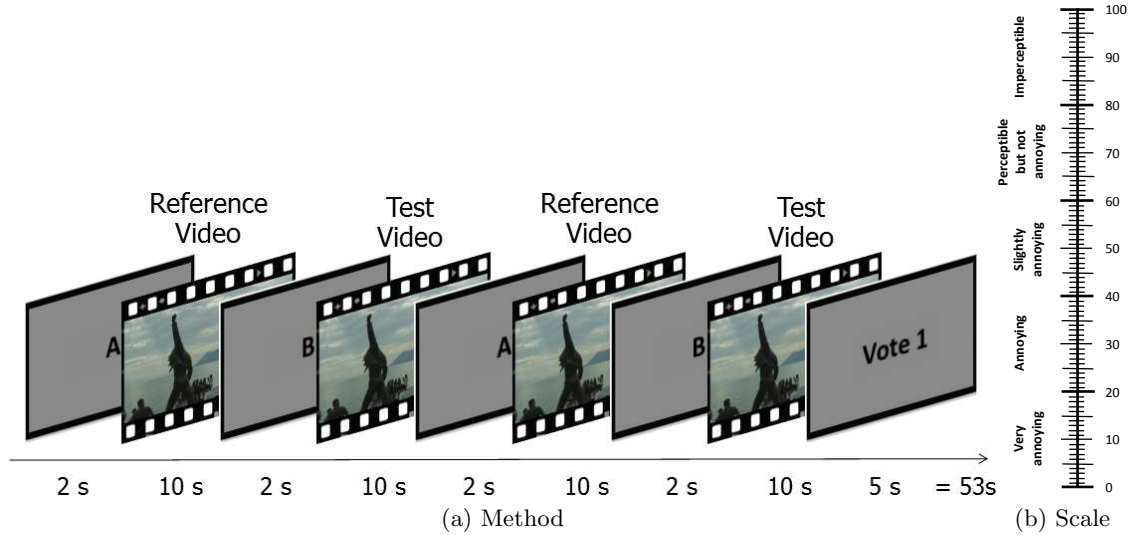


Figure 4: DSIS Variant II with continuous impairment scale.

the systematic error caused by the presentation order, the stimuli should be randomized in a way that the same content is never shown consecutively. To stabilize the rating of the observers, some dummy sequences can be included at the beginning of the test. Subjects are not told about the presence of dummy sequences and their scores are not included in the results. Additionally, one reference versus reference stimulus pair is usually included in the test to check the reliability of the subjects.

As shown in Figure 4(a), one stimuli presentation according to DSIS methodology, i.e., two consecutive presentations of two stimuli and rating time, theoretically takes 53 seconds. In total, 60 video stimuli in the evaluation campaign were evaluated considering all possible combination of used codecs, contents, and bit-rates. Therefore, the entire subjective evaluation was divided into three sessions. The first test session contained 23 presentations (i.e., 2 dummies video sequences + 1 reference vs. reference + 20 test video stimuli), corresponding to approximately 21 minutes. The remaining two test sessions contained 20 presentations (i.e., 20 test video stimuli), corresponding to about 18 minutes. However, as the player used in our experiment (Media Player Classic 64 bit) needs some time to load each video stimuli. The actual durations of all sessions were approximately 3 minutes longer.

The test was conducted within one day in four time slots per day. Each time slot was attended by six subjects, which were split into two groups of three subjects each. While one group was evaluating one session in the test room, the other group was resting in a separate room. Subjects assessing the test material were seated in three different positions (Left, Centre, and Right) with respect to the center of the monitor. A total of twenty four naive observers took part in the test campaign. Three of the observers were female and the age of the subjects ranged from 19 to 35 years old, with a median of 25.5 years old. Before the test itself, all participants were screened for correct visual acuity and color vision using Snellen and Ishiara charts, respectively. The training of the subjects of each group was conducted before the first test session, as a 10-minute training session, where oral instructions were provided to explain the task and a viewing session was performed to allow the subject to familiarize with the assessment procedure. The video sequences used as training samples had quality levels representative of the labels reported on the rating scales. The experimenter explained the meaning of each label reported on the scale and related them to the presented sample sequences.

To collect evaluation scores, subjects were provided with scoring sheets to enter their quality scores. The scores were then offline converted into electronic version. All the scores were converted by one operator and crosschecked by a second operator to identify and correct any eventual manual mistake.

3.4 Data processing

For the objective comparison of the encoders, the RD curves of combined luma (Y) and chroma (U, V) components were first used. Resulting PSNR_{YUV} values for each video content and bit-rate are calculated as an average of measurements per all frames. PSNR_{YUV} values for each frame, content, and bit-rate were calculated as a weighted sum of the PSNR values of each individual component,⁵ i.e., of PSNR_Y , PSNR_U , PSNR_V as follows:

$$\text{PSNR}_{YUV} = \frac{6 \times \text{PSNR}_Y + \text{PSNR}_U + \text{PSNR}_V}{8}. \quad (1)$$

On the other hand, the collected subjective raw scores were processed according to the following procedure. At first, the standard outlier detection¹⁴ was applied in each set of subjective scores assigned to each test sequence in order to remove subjects whose scores deviated strongly from the other scores in the same session. A subject was considered as an outlier, and thus all her/his scores were removed from the results of the session, if more than 20% of her/his scores over the session were outliers. In our experiments, none of the subjects was detected as an outlier for any of the test sessions.

To evaluate the perceived quality, the standard statistical indicators describing the score distribution across the subjects for each of the test conditions (combination of content, codec, and bit-rate) were computed. More specifically, the mean opinion score (MOS) and 95% confidence intervals (CI) using the Student's t-distribution were computed for each test condition.¹⁴

Finally, a multiple comparison analysis¹⁵ was performed, in order to identify the statistically significant differences among the MOS values obtained for different codecs and the similar bit-rate condition.

4. EXPERIMENTAL RESULTS

Left column of Figure 5 shows the PSNR RD curves corresponding to each test conditions (contents, codecs, and bit-rates). It can be clearly seen that from the estimated quality point of view, HEVC outperforms both coding algorithms AVC and VP9 for all contents and bit-rates. Furthermore, whereas AVC based encoding performs better for natural content, it loses by far for synthetic content in comparison to VP9 encoder.

Middle column of Figure 5 shows the RD plots with MOS and CI values for all contents, bit-rates, and codecs. Relatively small confidence intervals indicate high reliability of the results and rather small variation among the subjects. Having a closer look, the resulted plots show that HEVC outperforms both, its predecessor as well as its concurrent scheme VP9, for all the contents. Substantial quality improvements of HEVC coding algorithm in relation to AVC and VP9 are visible especially for lower bit-rates. Mutual comparison of AVC and VP9 brings credit to AVC especially for contents with higher TI (*Manege*, *Traffic*). On the other hand, VP9 outperforms AVC for synthetic content *Sintel2*.

Most of the applications require either fully or close to transparent quality of compressed video. Therefore, the performance analysis of the three codecs for operating points that are close to transparent (semi-transparent) or fully transparent in terms of perceived quality of compressed video sequences (i.e. above 75 in MOS scale) was performed. The actual bit-rate savings achieved by each codec, while maintaining the perceived quality level at transparent or semi-transparent level was computed across all natural and synthetic contents. We assume that the targeted transmission bit-rate for each codec and content corresponds to R5 (see Table 3). Then, the bit-rate savings of codec and content is computed as a ratio of corresponding R5 and a bit-rate where the perceived quality is equal 75 in MOS scale. It can be observed (see Figure 5 middle column), that while maintaining the perceived quality at transparent or semi-transparent level, HEVC allows a bit-rate savings of 68.9% and 71.9% in average for natural and synthetic content, respectively. On the other hand, AVC and VP9 encoders perform similarly taking into account an average results across all considered contents. More specifically, AVC outperforms VP9 for natural content, where it allows a bit-rate savings of 43% in comparison to only 24.3% bit-rate savings offered by VP9 to preserve the semi-transparent or fully transparent quality level. On the other hand, VP9 is more stable solution for synthetic content allowing a bit-rate reduction of 56.7%, whereas AVC can afford only 26.8% bit-rate savings to keep the transparent or semi-transparent quality level.

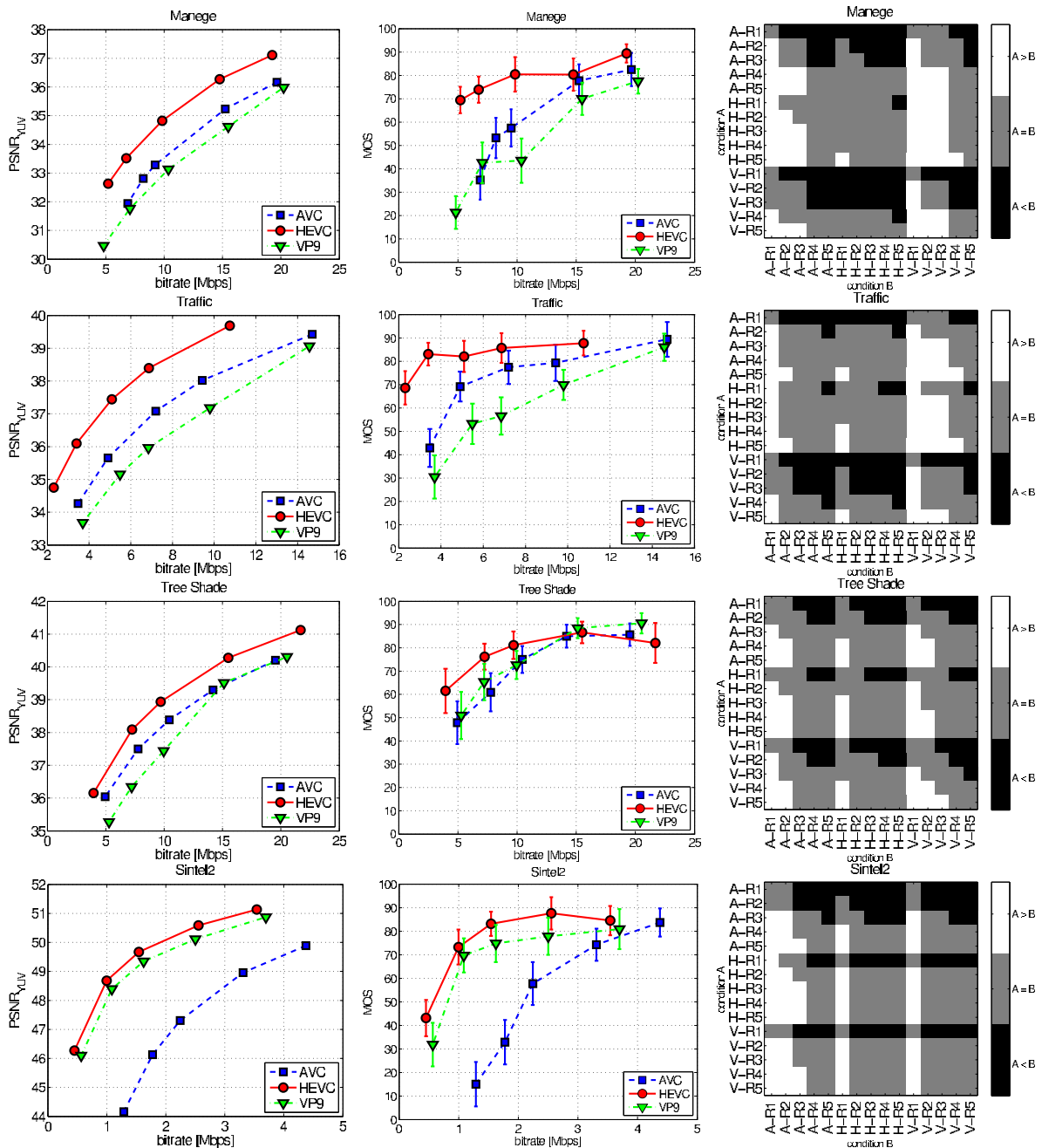


Figure 5: PSNR RD curves (left column) and subjective ratings showing mean opinion scores and confidence intervals (middle column) for different bit-rate and content separately. Right column presents a multiple comparison test for all possible combinations of codecs (A stands for AVC, H stands for HEVC, and V stands for VP9 coding algorithm) and bit-rates (R1 to R5), for each test content separately. In each plot, the color of each square shows the result of the significance test between the mean opinion scores related to the two test conditions reported in the corresponding column and row. A white (black) square indicates that the MOS corresponding to condition A is statistically significantly better (worse) than the MOS corresponding to condition B while a grey square indicates that the two MOSs are statistically not different.

Content	HEVC vs AVC		VP9 vs AVC		HEVC vs VP9	
	BD-PSNR	BD-MOS	BD-PSNR	BD-MOS	BD-PSNR	BD-MOS
Manege	-28.7%	-44.6%	10.6%	29.2%	-39.7%	-63.7%
Traffic	-37.2%	-57.5%	25.1%	61.0%	-49.8%	-74.7%
Tree Shade	-22.7%	-37.4%	18.9%	-8.2%	-33.7%	-31.9%
Sintel2	-69.9%	-70.9%	-60.9%	-61.7%	-19.0%	-27.5%
Average	-39.6%	-52.6%	-1.59%	5.1%	-35.6%	-49.4%

Table 5: Comparison of investigated coding algorithms in terms of bit-rate reduction for similar $PSNR_{YUV}$ and MOS. Negative values indicate actual bit-rate reduction.

To accurately analyze the performance of codecs and to evaluate whether the obtained results were significantly different when comparing one codec to another, a multiple comparison significance test has been applied to the data, for each combination of content and bit-rate separately.¹⁵ To identify all combinations of codecs and bit-rates, which resulted in statistically different mean opinion scores, a one-way ANOVA and multiple comparison tests were performed.¹⁵ Right column of Figure 5 shows the results comparing all possible conditions, i.e. codecs and bit-rates combinations, for each content separately. Comparing HEVC and AVC based encoding at similar bit-rates, the performance of HEVC is significantly better than AVC at three lower (R1-R3) bit-rates for *Manege* and *Sintel2*, and at lowest bit-rate (R1) for *Traffic*. Performance of HEVC and AVC for content *Tree Shade* is not significantly different at any of the bit-rates. Looking at results of VP9 vs AVC for natural content, it can be observed that the differences between perceived quality are not significant. On the other hand, VP9 outperforms AVC for synthetic content at three bit-rates (R1-R3). Perceived quality of contents *Manege* and *traffic* encoded by HEVC is significantly better at three bit-rates (R1-R3) in comparison to VP9 encoding. For the remaining two contents, HEVC and VP9 show similar performance.

The bit-rate reduction of one codec over another for a similar quality can be estimated using the Bjontegaard Delta PSNR (BD-PSNR).¹⁶ We used the combined $PSNR_{YUV}$ and bit-rate values as an input to the BD-PSNR measurement method, which allows to determine a single average difference in bit-rate that considers the tradeoffs between luma and chroma components fidelity.⁵ For more realistic comparison between individual coding algorithm, BD-MOS based on perceived quality can be defined and computed. BD-MOS uses MOS values instead of PSNR values to get a subjective coding gain for similar visual quality. For computation of BD-PSNR and BD-MOS, the SCENIC tool¹⁷ was used. Results for objective and subjective Bjontegaard Delta are shown in Table 5.

Based on estimated quality, and subjective scores, the average bit-rate reduction of the HEVC relative to AVC and VP9 coding schemes is 39.6% and 35.6%, and 52.6% and 49.4%, respectively. Nevertheless, taking into account only synthetic content, HEVC achieves 69.9% and 19.0% bit-rate reduction in comparison to AVC and VP9, respectively, which corresponds to bit-rate reduction of 70.9% and 27.5% based on subjective results. On the other hand, PSNR values for the natural content indicate that HEVC achieves total bit-rate savings of 22.7% to 37.2% in comparison to AVC and 33.7% to 49.8% in comparison to VP9. If we compare the average coding efficiency of VP9 versus AVC across investigated contents, it can be observed that VP9 can reduce the bit-rate by 1.59% in terms of PSNR and it requires 5.1% more bit-rate to achieve similar perceived quality. Taking into account only natural content considered in this study, the AVC can achieve 18.2% average bit-rate reduction based on PSNR and 27.3% average bit-rate reduction based on subjective scores in comparison to VP9. For Synthetic content, VP9 can reduce the bit-rate by approximately 61% to achieve similar objective and subjective quality in comparison to AVC.

In most cases of the natural content considered in our experiments, the actual coding gain of one codec to another is underestimated. more specifically, the actual perceived quality appears to be better and doesn't correspond to quality estimation performed by objective measures for a similar bit-rates. There is one exception for HEVC vs VP9 comparison for *Tree Shade* content where the perceived quality is well predicted by its estimation.

5. CONCLUSION

In this paper, a detailed description of the subjective quality evaluation test conducted to benchmark the performance of the emerging video compression specifications HEVC and VP9 for ultra high definition television broadcast scenario has been presented. The above mentioned compression algorithms have been compared to AVC anchor, the current state-of-the-art video coding standard, in the framework of four different UHD contents with various spatio-temporal characteristics. In total, four different UHD contents have been encoded at five bit-rates for each codec and subjective quality scores of 24 subjects related to each content, codec and quality have been collected. The obtained subjective scores show high consistency and allow an accurate comparison of the performance of the investigated codecs.

The test results show that HEVC offers improvements in compression performance when compared to VP9 and AVC, if one considers a wide range of bit-rates from low to high, corresponding to video with low to transparent quality. In this range of widest bit-rates and quality, objective based measurements shows that HEVC achieves average bit-rate savings of 39.6% versus AVC and 35.6% versus VP9. In the same wide range of bit-rates and quality, subjective scores show an average bit-rate reduction of HEVC by 52.6% in comparison to AVC and 49.4% in comparison to VP9. Similarly, comparison of coding efficiency between VP9 and AVC results slightly in favor of AVC (5.1%) in terms of subjective scores and slightly in favor on VP9 (1.59%) in terms of PSNR. Differences in bit-rates corresponding to close to transparent or transparent video are more nuanced and generally result in similar performance or better performance in favor of HEVC when compared to VP9 and AVC, especially for natural video content. Furthermore, VP9 seems particularly efficient and similar in performance to HEVC for synthetic video content.

The presented study focused on UHD TV broadcast scenarios, with a limited set of video content which could be used for scientific research and published. However, a more complex comparison of mentioned coding schemes should also include the Internet streaming scenarios such as those in video sharing platforms, and a wider set of content. This more complex approach will be investigated in the future as an extension of what is presented in this paper.

ACKNOWLEDGMENTS

This work has been performed in the framework of the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services - QUALINET, FP7 EC EUROSTAR funded Project - Transcoders Of the Future TeleVision(TOFuTV), and EC funded Network of Excellence VideoSense.

REFERENCES

- [1] ISO, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 2: High efficiency video coding," Tech. Rep. ISO/IEC 23008-2:2013, ISO/IEC (2013).
- [2] ISO, "Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding," Tech. Rep. ISO/IEC 14496-10:2012, ISO/IEC (2012).
- [3] Bankoski, J., Wilkins, P., and Xu, Y., "Technical overview of vp8, an open source video codec for the web," in [*Multimedia and Expo (ICME), 2011 IEEE International Conference on*], 1–6 (2011).
- [4] Bankoski, J., Bultje, R. S., Grange, A., Gu, Q., Han, J., Koleszar, J., Mukherjee, D., Wilkins, P., and Xu, Y., "Towards a next generation open-source video codec," in [*IS&T/SPIE Electronic Imaging*], 866606–866606, International Society for Optics and Photonics (2013).
- [5] Ohm, J., Sullivan, G. J., Schwarz, H., Tan, T. K., and Wiegand, T., "Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC)," *Circuits and Systems for Video Technology, IEEE Transactions on* **22**(12), 1669–1684 (2012).
- [6] Hanhart, P., Rerabek, M., De Simone, F., and Ebrahimi, T., "Subjective quality evaluation of the upcoming HEVC video compression standard," in [*SPIE Optical Engineering+ Applications*], 84990V–84990V, International Society for Optics and Photonics (2012).
- [7] Grois, D., Marpe, D., Mulayoff, A., Itzhaky, B., and Hadar, O., "Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders," in [*PCS*], 394–397 (2013).

- [8] Wiegand, T., Sullivan, G. J., Bjontegaard, G., and Luthra, A., “Overview of the H. 264/AVC video coding standard,” *Circuits and Systems for Video Technology, IEEE Transactions on* **13**(7), 560–576 (2003).
- [9] Sullivan, G. J., Ohm, J., Han, W.-J., and Wiegand, T., “Overview of the high efficiency video coding (HEVC) standard,” *Circuits and Systems for Video Technology, IEEE Transactions on* **22**(12), 1649–1668 (2012).
- [10] Song, L., Tang, X., Zhang, W., Yang, X., and Xia, P., “The SJTU 4K video sequence dataset,” in [*QoMEX*], 34–35 (2013).
- [11] ITU-R, “P.910: Subjective video quality assessment methods for multimedia applications,” Tech. Rep. P.910, ITU-R (1992).
- [12] ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures.” International Telecommunication Union (January 2012).
- [13] ITU-R BT.2022, “General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays.” International Telecommunication Union (August 2012).
- [14] De Simone, F., Goldmann, L., Lee, J.-S., and Ebrahimi, T., “Towards high efficiency video coding: Subjective evaluation of potential coding technologies,” *Journal of Visual Communication and Image Representation* **22**(8), 734 – 748 (2011).
- [15] Snedecor, G. W. and Cochran, W. G., [*Statistical Methods*], Iowa State University Press (1989).
- [16] Bjontegaard, G., “Improvements of the BD-PSNR model,” Tech. Rep. VCEG-AI11, ITU-T SG16/Q6, Berlin, Germany (July 2008).
- [17] Hanhart, P. and Ebrahimi, T., “Calculation of average coding efficiency based on subjective quality scores,” *Journal of Visual Communication and Image Representation* **25**(3), 555–564 (2014).