Subjective quality evaluation of the upcoming HEVC video compression standard

Philippe Hanhart, Martin Rerabek, Francesca De Simone, and Touradj Ebrahimi

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

ABSTRACT

High Efficiency Video Coding (HEVC) is the latest attempt by ISO/MPEG and ITU-T/VCEG to define the next generation compression standard beyond H.264/MPEG-4 Part 10 AVC. One of the major goals of HEVC is to provide efficient compression for resolutions beyond HDTV. However, the subjective evaluations that led to the selection of technologies were bound to HDTV resolution. Moreover, performance evaluation metrics to report efficiency results of this standard are mainly based on PSNR, especially for resolutions beyond HDTV. This paper provides subjective evaluation results to assess the performance of the current HEVC codec for resolutions beyond HDTV.

Keywords: subjective quality assessment, High Efficiency Video Coding (HEVC), Joint Collaborative Team on Video Coding (JCT-VC), H.264/MPEG-4 Part 10 AVC

1. INTRODUCTION

The current trend in video consumption clearly shows that the already large quantity of video material distributed over broadcast channels, digital networks, and packaged media is going to increase in the coming years. As an effect of the growing popularity, the users' demand for increased resolution and higher quality is driving the efforts of the technological development. From this point of view, the evolution of video acquisition and display technologies is much faster than that of network capabilities. Thus, a clear need for a new video coding standard with higher efficiency when compared to the current state-of-the-art H.264/MPEG-4 Part 10,[1] referred to as AVC (Advanced Video Coding) in the rest of the paper, has been identified.

In order to develop the next-generation video coding standard, a 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)*, has been created. The JCT-VC standardization effort is being referred to as High Efficiency Video Coding (HEVC). The new standard targets a wide variety of applications such as mobile TV, home cinema, and ultra high definition television (UHDTV). It will support next-generation acquisition and display devices featuring progressive scanned video with higher frame rates and resolutions (from WVGA to HDTV and UHDTV), as well as improved picture quality in terms of noise level, color gamut, and dynamic range. HEVC aims at a substantially improved coding efficiency compared to the AVC High Profile, i.e., reducing the bit rate requirements by half while keeping comparable image quality, probably at the expense of increased computational complexity. Depending on the application scenario, a trade-off between computational complexity, compression ratio, robustness to errors, and processing delay should be supported.

A Joint Call for Proposals (CfP) on Video Compression Technology[2] was issued by JCT-VC in January 2010. A total of 27 proposals were evaluated with respect to two AVC anchors in the largest subjective video quality testing effort ever conducted.[3,4] All proposals used a coding architecture conceptually similar to AVC, containing the following basic elements: (a) Block-based coding (b) Variable block sizes (c) Block motion compensation (d) Fractional-pel motion vectors (e) Spatial intra prediction (f) Spatial transform of residual difference

---

Further author information: (Send correspondence to Philippe Hanhart)

Philippe Hanhart: E-mail: philippe.hanhart@epfl.ch

Martin Rerabek: E-mail: martin.rerabek@epfl.ch

Francesca De Simone: E-mail: francesca.desimone@epfl.ch

Touradj Ebrahimi: E-mail: touradj.ebrahimi@epfl.ch

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

Dataset	Video	Resolution	Framerate
Test	PeopleOnStreet	3840x2160	30
	Traffic	3840x2048	30
	Sintel2	3840x1744	24
Training	Sintel39	3840x1744	24

Table 1: Dataset

(g) Integer-based transform designs (h) Arithmetic or VLC-based entropy coding (i) In-loop filtering. However, the individual coding tools differed a lot between the individual proposals. Key elements of some of the best proposals were combined to develop an initial Test Model, as a starting point for the definition of the new standard.[4] The initial Test Model was refined over the next JCT-VC meetings and, in January 2011, an official Test Model, named HEVC Test Model (HM), was publicly released.[5] The HM software integrates the latest developments that have been validated within the JCT-VC group and a new version is available at each JCT-VC meeting cycle.

It is expected that HEVC achieves double the compression efficiency of AVC, at the expense of a significant increase in computational complexity. The compression efficiency of different codecs can be reliably compared only by means of subjective tests, carried out according to common evaluation methodologies defined by experts. Therefore, the responses to the CfP were evaluated during a formal subjective test campaign[3,6] and informal subjective tests are still carried out during the development of the standard to assess the improvements of the integrated coding tools.[7,8] It is expected that HEVC achieves even better compression efficiency for resolutions beyond HDTV, especially due to increased prediction flexibility and a wider range of block sizes. However, to the best of our knowledge, no subjective evaluation has been performed so far on resolutions higher than HDTV, mostly because of hardware limitations and the lack of high quality uncompressed content.

In this paper, we present the details and the results of a subjective quality evaluation performed on a professional high-performance 4K/Quad Full High Definition (QFHD) LCD reference monitor to benchmark the performance of HEVC and AVC on 4K/QFHD video content.

The paper is structured as follows: the laboratory environment, where the subjective test took place, is described in Section 3. The test material and the adopted test methodology are detailed in Section 2, while the statistical analysis of the collected subjective data and the results are presented in Section 4 and Section 6, respectively. Finally, concluding remarks are drawn in Section 7.

2. DATASET AND TEST METHODOLOGY

2.1 Dataset

The availability of high quality 4K uncompressed video data free of use for research purpose is very limited. Only two contents are available to the JCT-VC group: PeopleOnStreet and Traffic. To cover a wider application scenario, synthetic content from the Sintel movie† was included. 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 four contents, one for the training and three for the test, with different visual characteristics, resolutions, and frame rates, as shown in Table 1. The first frame of each content is shown in Figure 1. Figure 2 shows the spatial information (SI) and temporal information (TI) indexes on the luminance component of each content, as indicated in.[9] It is observed that Sintel2 and Sintel39 have large TI values, while Traffic shows a small TI index. Since the Traffic sequence is five seconds long only, it was decided to clip all contents to five seconds to maintain consistency during the test between the different contents. All test sequences were stored as raw video files, progressively scanned, with YUV 4:2:0 color sampling, and 8 bits per sample.

---

†Sintel is a computer animated movie produced by the Blender Institute using the open source computer graphics software Blender and released under the Creative Commons Attribution license. http://www.sintel.org

(a) PeopleOnStreet

(b) Traffic

(c) Sintel2

(d) Sintel39

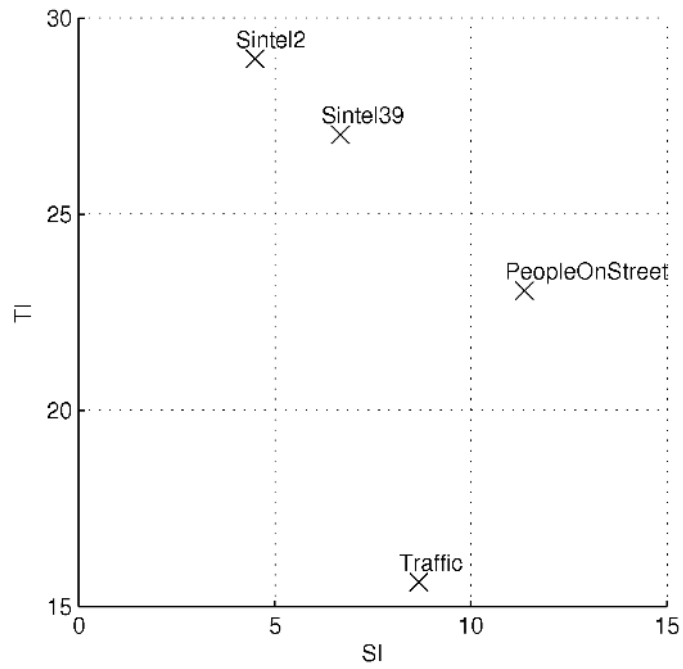Figure 1: Sample frames of the individual contents considered in the subjective test.

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

Codec	AVC	HEVC
Encoder	JM 18.3	HM 6.1.1
Profile	High 5.1	Main
Reference Frames	4	4
R/D Optimization	On	On
Motion Estimation	EPZS	EPZS
Weighted Prediction	On	-
Search Range	128	64
Group of Pictures	8	8
Hierarchical Encoding	On	On
Temporal Levels	4	4
Intra Period	1s	1s
Deblocking	On	On
Rate Control	Off	-
8x8 Transform	On	-
Adaptive Loop Filter	-	Off
Coding Unit size / depth	-	64 / 4
Transform Unit size min / max	-	4 / 32

Table 2: Selected encoder settings for AVC and HEVC.

The video sequences were compressed with AVC and HEVC using JM 18.3 and HM 6.1.1, respectively. The Random Access (RA) configuration was selected for this study 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 a 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 such that similarity was ensured between the two codecs to avoid penalization. For example, *BLevel0MoreRef* and *BIdenticalList* were set in the JM configuration file. More details on the configurations can be found in Table 2.

For each content and codec, five different bit rates were selected. Due to the different spatio-temporal characteristics of the contents and the presence of both natural and synthetic content, it was decided to select the targeted bit rates for each content separately. Since no Rate Control is implemented in HM 6.1.1, fixed QPs were used. Typical QPs for AVC are in the range of 25 to 37. First, a few sequences were compressed for each content using this range, keeping in mind the $\sim$ 12.5% per QP rule (there is approximately a 12.5% bit rate reduction for every increase in QP), and trying to map the QPs of the HM to those of the JM. To be realistic, it was decided to set the upper bit rate limit to 20 Mbps. Then, an expert screening session was conducted to select the lower and upper bounds for each content separately, keeping in mind the standard QP range and targeting realistic bit rates, to try to cover the full quality scale. Finally, the targeting 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 and HEVC degradations with similar strengths were selected. The complete sets of targeted and actual bit rates are shown in Table 3 and Table 4, respectively. The QPs used to encode these sequences are specified in Table 5 and the resulting PSNR plots are shown in Figure 3.

Content	Codec	R1	R2	R3	R4	R5
PeopleOnStreet	AVC	5.000	7.000	10.000	14.000	20.000
	HEVC	5.000	7.000	10.000	14.000	20.000
Traffic	AVC	3.500	5.000	7.000	10.000	14.000
	HEVC	2.500	3.500	5.000	7.000	10.000
Sintel2	AVC	1.200	1.600	2.000	2.500	3.500
	HEVC	0.768	1.200	1.600	2.000	2.500

Table 3: Targeted bit rates (Mbps).

Content	Codec	R1	R2	R3	R4	R5
PeopleOnStreet	AVC	4.743	6.799	9.454	14.561	20.745
	HEVC	4.889	6.960	9.833	13.871	20.278
Traffic	AVC	3.490	4.914	7.208	9.429	14.717
	HEVC	2.277	3.346	4.997	6.720	10.474
Sintel2	AVC	1.205	1.571	1.935	2.389	3.455
	HEVC	0.705	1.204	1.616	1.903	2.674

Table 4: Actual bit rates (Mbps).

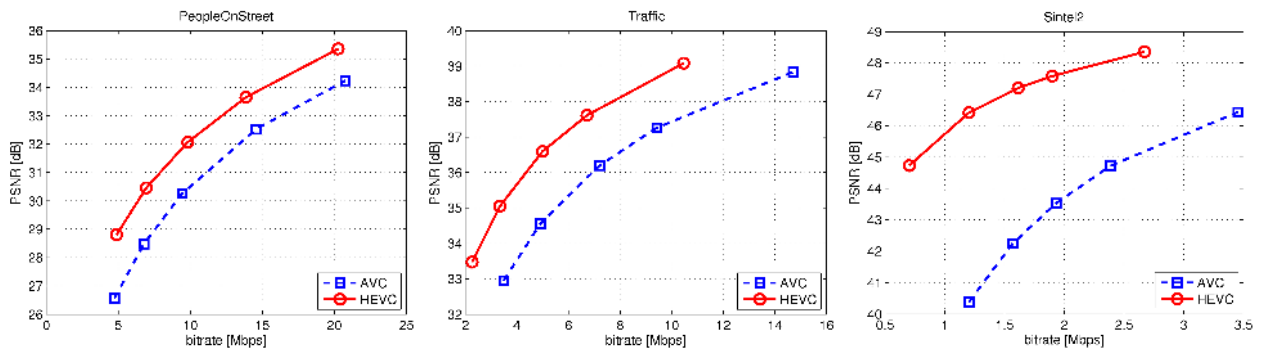| Content	Codec	R1	R2	R3	R4	R5
PeopleOnStreet	AVC	44	41	38	34	31
	HEVC	42	39	36	33	30
Traffic	AVC	37	34	31	29	26
	HEVC	38	35	32	30	27
Sintel2	AVC	35	32	30	28	25
	HEVC	32	28	26	25	23

Table 5: Quantization Parameters (QPs).
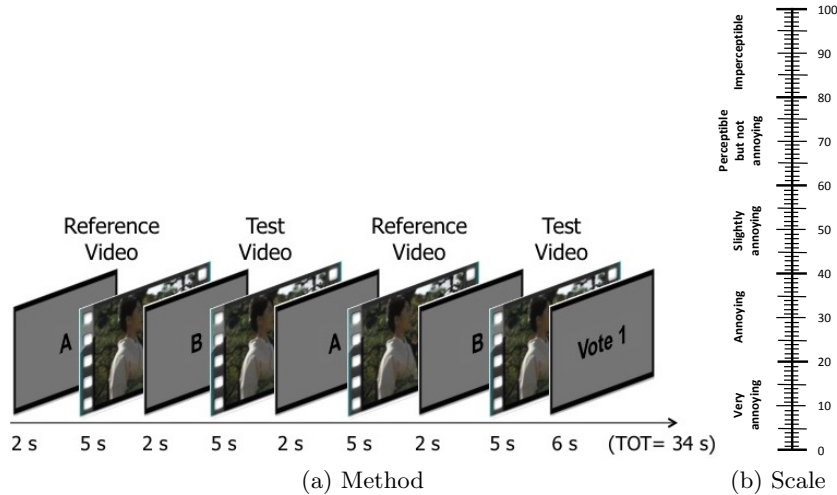

Figure 3: PSNR versus bit rate.

Figure 4: DSIS Variant II with continuous impairment scale.

2.2 Test methodology

Since the test sequences are only 5 seconds long and subjects are not used to watch ultra high definition television, the Double Stimulus Impairment Scale (DSIS) method,[10] Variant II, with a continuous impairment scale was chosen to perform the subjective quality assessment experiments.

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, having the best expected quality, as stimulus A and she/he is asked to rate the level of annoyance of the visual defects that is observed in stimulus B. The used rating scale is shown in Figure 4(b).

2.3 Sessions plan

In order to retain the concentration of the subjects, a subjective video quality test session should not last more than 30 minutes.[10] Since the evaluation task requires a lot of attention due to the short sequences duration, it was decided to split the test in sessions that are no longer than 15 minutes each, followed by a resting phase. For the same reason, it is preferable to alternate as many different contents as possible in the same session. Furthermore, to avoid a possible effect of the presentation order, the stimuli are randomized in a way that the same content is never shown consecutively. Also, some dummy sequences, whose scores are not included in the results but the observer is not told about, are included at the beginning of the first session to stabilize observers' rating after training. Additionally, a reference versus reference stimulus pair can be included to check the reliability of the subjects. We decided to include 2 dummy presentations and 1 reference vs reference pair at the beginning of the first session.

As shown in Figure 4, one DSIS Variant II presentation, i.e., two consecutive presentations of two stimuli and rating time, theoretically takes 34 seconds. However, the video player used in our experiment takes about 4 seconds to load a video file. To reduce the waiting time, we concatenated the 'A' message, reference sequence, 'B' message, and test sequence into a single avi file for each test condition. Thus, only 3 loading times occurred during one complete DSIS presentation, one before each consecutive presentation and one before the voting message. Therefore, one DSIS Variant II presentation took about 46 seconds. We had to evaluate a total of 30 test sequences (i.e., 2 codecs × 3 contents × 5 bit rates), thus it was decided to split the test in two sessions. The first test session contained 18 presentations (i.e., 2 dummies + 1 ref vs ref + 15 stimuli), corresponding to about 14 minutes. The second test session contained 15 presentations (i.e., 15 stimuli), corresponding to about 11.5 minutes.

Category	Model
Motherboard	Intel DX58SO2, Chipset Intel X58 with ICH 10R
Processor	Intel Core i7 980X Extreme
Graphics	ATI Radeon Fire Pro V8800
RAM	Kingston Memory 3x4 GB PC3-10700
SDD (Playback)	OCZ RevoDrive 3 X2 SSD 480GB, max read: 1500MB/s, max write 1250 MB/s
HDD (Storage)	Western Digital 2x2 TB
Operating system	Windows 7 Enterprise 64 bit
Video player	Media Player Classic 64 bit

Table 6: Server configuration with hardware and software details.

The test was planned over 2 days, with 3 time slots per day. Each time slot was attended by 6 subjects, which were split into two groups of 3 subjects each. While one group was evaluating one session in the test room, the other group was resting in a separate room. A total of 36 naive people took part in the test campaign. 30% of the observers were female and the age of the subjects ranged from 20 to 61 years old, with a median of 25 years old. All participants were screened for correct visual acuity and color vision using Snellen charts and Ishiara charts, respectively. A picture of the screening area in shown in Figure 5(a).

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. TEST EQUIPMENT AND ENVIRONMENT

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. The test has to be carried out according to precise methodologies and in a controlled test environment to produce reliable and reproducible results, avoiding involuntary influence of external factors.[10]

Natural playback in native spatial and temporal resolutions of raw 4K/QFHD video sequences at 30 fps requires specific hardware. Particularly, reading and displaying in real time YUV 4:2:0 color subsampled QFHD (3840 × 2160 pixels) video sequences at 30 fps requires a data rate of 373.25 MB/s. Since the typical reading speed of current Hard Disk Drives (HDD) is below 160 MB/s, a hardware solution based on Solid State Drives (SSD) was adopted. The details of the video server and the software used to display the video sequences are listed in Table 6.

To display 4K/QFHD content, a 56-inch professional high-performance 4K/QFHD LCD reference monitor Sony Trimaster SRM-L560 [‡] was used. The monitor consists of four Full HD panels. The panels are driven by four display ports and mutually synchronized by the graphic board of the video server to prevent any tearing effect. This monitor can operate in tree different modes (4K/QFHD, Quad View, and 2K/HD Zoom), while only the first one is available when DVI inputs are used.

To assure the reproducibility of results by avoiding involuntary influence of external factors, the laboratory for subjective video quality assessment was set up according to.[10] 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

---

[‡] http://pro.sony.com/bbsccms/assets/files/cat/mondisp/brochures/di0195_srm1560.pdf

(a) Screening area   (b) Testing area

Figure 5: MMSPG subjective visual quality test laboratory, compliant with ITU recommendation.[10]

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 20 lux and the ambient black level was 0.2 cd/m^2. The test area was controlled by an indoor video security system to keep track of all the test activities and of possible unexpected events, which could influence the test results. A picture of the MMSPG test environment where assessments took place is shown in Figure 5(b).

The experiments involved three subjects assessing the test material, seated in three different positions (Left, Centre, and Right) with respect to the center of the monitor. The distance of the subjects from the monitor was approximately equal to 3.5 times the height of the display.

4. DATA PROCESSING

4.1 Outlier detection

To detect and remove subjects whose scores appear to deviate strongly from the other scores in a session, outlier detection was performed. The outlier detection was applied to the set of results obtained from the 36 subjects.

In each set of scores assigned to a test sequence, a score by subject j and test condition i, s_{ij}, was considered as outlier if $s_{ij} > q_3 + 1.5(q_3 - q_1) \vee s_{ij} < q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles of the scores distribution for test condition i, respectively.[6] This range corresponds to approximately ± 2.7 the standard deviation or 99.3% coverage if the data is normally distributed. 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.[6] In this study, no outlier subjects were detected.

4.2 Statistical measures

Statistical measures were computed to describe the score distribution across the subjects for each of the test conditions (combination of content, codec, and bit rate).

The mean opinion score (MOS) is computed as:

$$MOS_i = \frac{\sum_{j=1}^{N} s_{ij}}{N} \tag{1}$$

where N is the number of valid subjects and s_{ij} is the score by subject j for the test condition i.

The relationship between the estimated mean values based on a sample of the population (i.e., the subjects who took part in our experiments) and the true mean values of the entire population is given by the confidence interval of the estimated mean. The $100 \times (1 - \alpha)\%$ confidence intervals (CI) for MOSs were computed using the Students t-distribution, as follows:

$$CI_i = t(1 - \alpha/2, N) \cdot \frac{\sigma_i}{\sqrt{N}} \tag{2}$$

where $t(1 - \alpha/2, N)$ is the t-value corresponding to a two-tailed t-Student distribution with $N - 1$ degrees of freedom and a desired significance level α (equal to 1-degree of confidence). N corresponds to the number of valid subjects, and σ_i is the standard deviation of a single test condition i across the subjects j. The confidence intervals were computed for an α equal to 0.05, which corresponds to a degree of significance of 95%.

5. PERFORMANCE INDEXES

The results of the subjective tests can be used as ground truth to evaluate how well PSNR estimates perceived quality. As compliant to the standard procedure for evaluating the performance of objective metrics,[11] the following properties of the PSNR estimation of MOSs are considered in this study: accuracy, monotonicity, and consistency.

First, a logistic function is fitted in a least squares sense to the [PSNR, MOS] data set for each content separately. The logistic function takes into account non-linearities and saturation effect of the human visual system. The logistic function is of the form:

$$\mathrm{MOS}_p(\mathrm{PSNR}) = \frac{a}{1 + \exp\left[-b\left(\mathrm{PSNR} - c\right)\right]}$$

Then, the Pearson linear correlation coefficient (PCC) and the root-mean-square error (RMSE) are computed between MOS_p and MOS to estimate the accuracy of the PSNR. To estimate monotonicity and consistency, the Spearman rank order correlation coefficient (SCC) and the outlier ratio (OR), are computed between MOS_p and MOS, respectively.

The root-mean-square error (RMSE) and the outlier ratio (OR) are defined as follow:

$$\mathrm{RMSE} = \sqrt{\frac{1}{(M - D)} \sum_{i=1}^{M} (\mathrm{MOS}_i - \mathrm{MOS}_{pi})^2}$$

$$\mathrm{OR} = \frac{\text{total number of outliers}}{M}$$

$$\text{outlier: point for which } |\mathrm{MOS}_i - \mathrm{MOS}_{pi}| > 2\sigma_i$$

where M is the total number of points, D is the degree of freedom for the curve fitting (logistic: $D = 3$), and σ_i is the standard deviation corresponding to MOS_i.

6. RESULTS AND DISCUSSION

Figure 6 shows resulting MOS and CI plots for the different contents. As it can be seen from the small confidence intervals, the results are reliable and the variations between the subjects are rather small. The results show that, especially for lower bit rates, the performance of HEVC exhibits a substantial quality improvement compared to AVC.

Traffic is relatively easy to encode since it has a small TI index. Therefore, bit rates as low as 5 Mbps and 2 Mbps for AVC and HEVC, respectively, are evaluated as transparent. In this case, a 60% bit rate reduction can easily be achieved. PeopleOnStreet is more challenging since it has higher SI and TI indexes, but mostly because artifacts are more visible in the upper left corner due to higher sensitivity of the human visual system
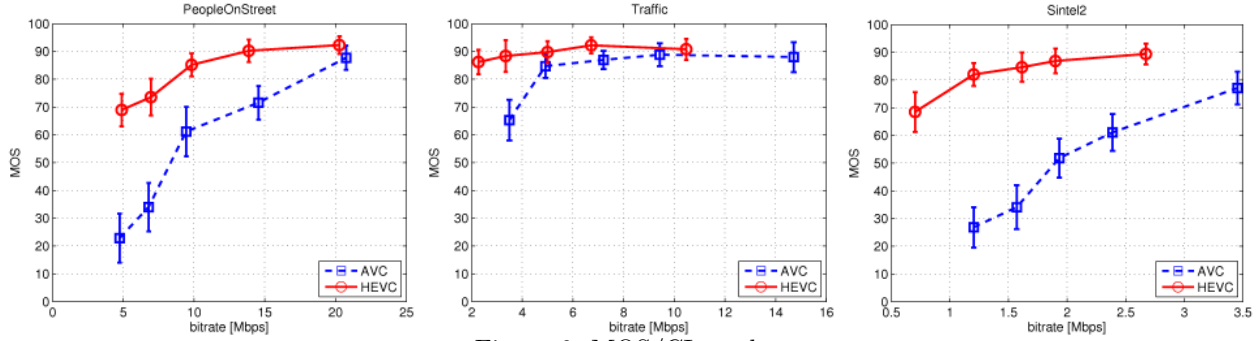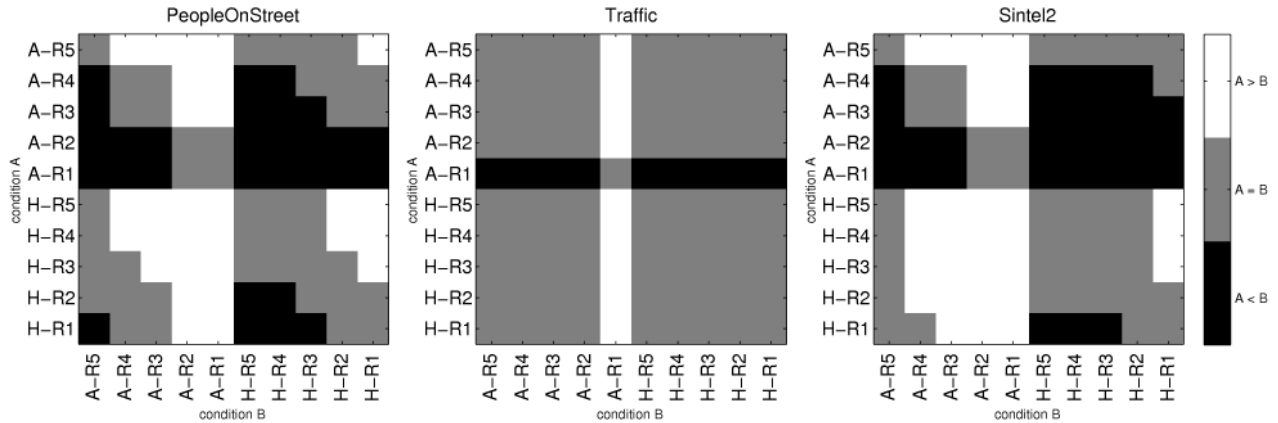
Figure 6: MOS/CI results.

Figure 7: Results of the multiple comparison test for the different test conditions, i.e., combination of codec (A stands for AVC and H stands for HEVC) and bit rate (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.

in low intensity areas (Weber law). For this content, blockiness was perceived in AVC encoded sequences while the content was smoothed out in HEVC encoded sequences, which is less annoying. For the synthetic content, HEVC exhibits a significant improvement over AVC and very low bit rates can be achieved due to the absence of noise in the original content. A bit rate as low as 1.2 Mbps is perceived as transparent with HEVC while the same bit rate for AVC is evaluated as annoying. For this content, a 70% bit rate reduction is achieved.

To accurately analyze the performance of HEVC and evaluate whether the obtained results were significantly different from those obtained with AVC, a multiple comparison significance procedure has been applied to the data, for each combination of content and bit rate separately.[12]

To identify the test conditions that resulted in statistically different mean opinion scores, a one-way ANOVA and multiple comparison tests were performed, considering as treatment the combination of codec and bit rate.[12] The results comparing all the possible pairs of treatments are shown in Figure 7, for each content separately. Comparing the two codecs at similar bit rates, HEVC outperforms AVC for 4 bit rates out of 4 for Sintel2 (1.2, 1.6, 2, and 2.5 Mbps) and for 4 bit rates out of 5 for PeopleOnStreet (5, 7, 10, and 14 Mbps), while only for 1 bit rate out of 4 for Traffic (3.5 Mbps). For the remaining bit rates, the codecs show the same performance.

A two-way ANOVA,[12] considering the codec and the bit rate as two separate treatments, has also been performed, resulting in a significant codec effect and significant bit rate effect, but irrelevant interaction effect.

Scatter plots of subjective versus objective results are presented in Figure 8 for each content separately. It can be noticed that the results for Sintel2 are in the linear phase, where an increase in PSNR induces an increase
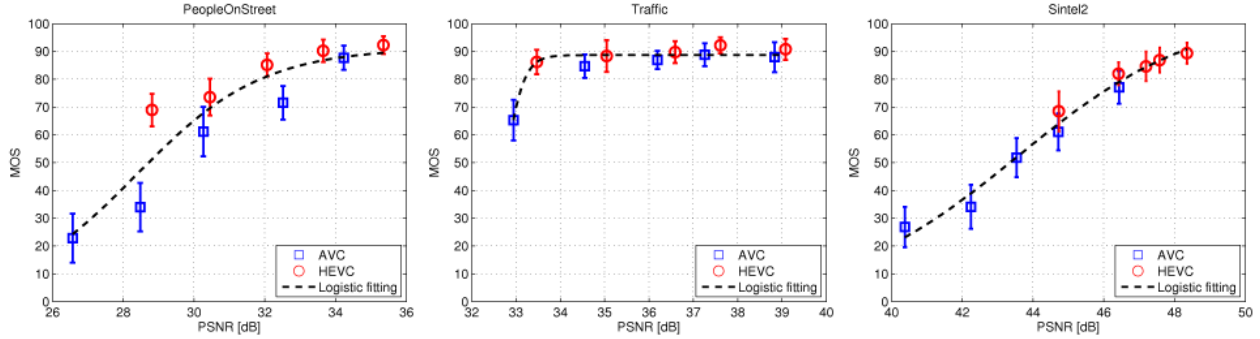
Figure 8: Subjective versus objective results.

Content	PCC	SCC	RMSE	OR
PeopleOnStreet	0.9266	0.9394	10.1034	0
Traffic	0.9640	0.8061	2.3004	0
Sintel2	0.9912	0.9879	3.3665	0
Average	0.9606	0.9111	5.2568	0

Table 7: Accuracy, monotonicity, and consistency indexes.

in perceived quality, while the results for Traffic are mostly in the saturation phase, where no significant gain in quality is perceived for an increase in PSNR. In both cases, the correlation between predicted and perceived quality is high. For PeopleOnStreet, most results are on the upper part of the linear phase and a few are on the saturation phase. The data points show a clear difference in terms of perceived quality between AVC and HEVC for a similar PSNR. The accuracy, monotonicity, and consistency indexes of the objective metric, as defined in Section 5, are reported for each content separately in Table 7. The resulting values confirm the graphical analysis. In general, PSNR is highly correlated with perceived quality, as long as the saturation limits are considered.

The bit rate reduction of HEVC over AVC for a similar quality can be estimated using the Bjontegaard Delta PSNR (BD-PSNR).[13] Since the original model uses 4 rate points and associated PSNR values, we computed the BD-PSNR for each ordered combination of 4 AVC and 4 HEVC bit rates and averaged the results across the different combinations. BD-PSNR gives an objective gain. Similarly, BD-MOS can be defined and computed using MOS values instead of PSNR values. BD-MOS gives a subjective gain. BD-MOS is more realistic since it is based on perceived quality while BD-PSNR is based on estimated quality. Results are shown in Table 8. It can be noticed that BD-PSNR under estimates the actual bit rate reduction, especially for Traffic. For this content, the difference is due to the saturation effect in perceived quality, which is not captured by PSNR. For Sintel2, the values are very similar since the relation between MOS and PSNR is almost linear for the considered bit rates. In the case of PeopleOnStreet, BD-PSNR also under estimates the actual gain because PSNR does not fully capture the difference between AVC and HEVC artifacts.

Content	Objectif BD-PSNR	Subjectif BD-MOS
PeopleOnStreet	27.5%	50.8%
Traffic	37.7%	74.0%
Sintel2	68.0%	74.7%
Average	44.4%	66.5%

Table 8: Bit rate reduction of HEVC over AVC for similar quality.

7. CONCLUSION

In this paper, a detailed description of the subjective quality evaluation performed to benchmark the performance of the upcoming HEVC video compression standard on ultra high definition television has been presented.

The next generation compression standard has been evaluated in comparison to the current standard, namely AVC, by means of subjective evaluation performed on resolution beyond HDTV. The evaluation was performed on three contents with different spatio-temporal characteristics, including both natural and synthetic contents, encoded at five different bit rates for each codec and content. Subjective quality scores related to a total of 30 test stimuli have been collected. The obtained results show high consistency and allow an accurate comparison of the performance of the two codecs.

The test results clearly exhibited a substantial improvement in compression performance, as compared to AVC. In most cases, a significant difference is observed between HEVC and AVC for a similar bit rate. For the natural contents considered in this study, a bit rate reduction ranging from 51 to 74% can be achieved based on subjective results while the predicted reduction based on PSNR values was only between 28 and 38%. This difference is mostly due to the fact that PSNR doesn't take into account the saturation effect of the human visual system. PSNR also doesn't capture the full nature of the artifacts: AVC compressed sequences exhibit blockiness while HEVC compression tends to smooth out the content, which is less annoying. For the synthetic content considered in this study, a 75% bit rate reduction can be achieved based on subjective results while the predicted reduction based on PSNR values was 68%.

According to JCT-VC, double the compression efficiency of AVC is expected to be achieved. This paper confirms that a significantly higher compression performance can be achieved on resolutions beyond HDTV, manly thanks to increased prediction flexibility and a wider range of block sizes. As ultra high definition television has recently been demonstrated to be the future of television, the upcoming HEVC video compression standard seems to be one of the key elements towards a wide deployment of 4K and 8K resolutions.

ACKNOWLEDGMENTS

REFERENCES

[1] ISO, "Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding," Tech. Rep. ISO/IEC 14496-10:2005, ISO/IEC (2005).

[2] ISO, "Joint Call for Proposals on Video Compression Technology," Tech. Rep. N11113, ISO/IEC JTC1/SC29/WG11 ITU-T Q6/16 Visual Coding, Kyoto, JP (Jan 2010).

[3] ISO, "Report of subjective testing of responses to Joint Call for Proposals (CfP) on video coding technology for High Efficiency Video Coding (HEVC)," Tech. Rep. M24092, ISO/IEC JTC1/SC29/WG11, Dresden, Germany (April 2010).

[4] Sullivan, G. J. and Ohm, J.-R., "Recent developments in standardization of high efficiency video coding (HEVC)," *Applications of Digital Image Processing XXXIII* **7798**(1), 77980V, SPIE (2010).

[5] ISO, "JCT-VC AHG report: Software development and HM software technical evaluation," Tech. Rep. M19119, ISO/IEC JTC1/SC29/WG11, Daegu, Korea (January 2011).

[6] 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).

[7] ISO, "JCT-VC AHG report: HM subjective quality investigation (AHG22)," Tech. Rep. M23863, ISO/IEC JTC1/SC29/WG11, San José, USA (February 2012).

[8] ISO, "[AHG8] Objective and subjective evaluation of HM5.0," Tech. Rep. M22988, ISO/IEC JTC1/SC29/WG11, San José, USA (February 2012).

[9] ITU-R, "P.910: Subjective video quality assessment methods for multimedia applications," Tech. Rep. P.910, ITU-R (1992).

[10] ITU-R, "BT.500-11: Methodology for the subjective assessment of the quality of television pictures," Tech. Rep. BT.500-11, ITU-R (2002).

[11] ITU-T, "Objective perceptual assessment of video quality: Full reference television." ITU-T Telecommunication Standardization Bureau (2004).

[12] Snedecor, G. W. and Cochran, W. G., [*Statistical Methods*], Iowa State University, Press (1989).

[13] Bjontegaard, G., "Improvements of the BD-PSNR model," Tech. Rep. VCEG-AI11, ITU-T SG16/Q6, Berlin, Germany (July 2008).