Joint Video Experts Team (JVET)
of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29
20th Meeting, by teleconference, 7–16 Oct. 2020

Document: JVET-T2020

Title: **VVC Verification Test Report for Ultra High Definition (UHD) Standard Dynamic Range (SDR) Video Content**

Status: Output document approved by JVET

Purpose: Report

Author(s) or Contact(s): Mathias Wien Email: wien@lfb.rwth-aachen.de
Vittorio Baroncini baroncini@gmx.com

Source: Verification Test Coordinators

Executive Summary

This document reports verification test results comparing VVC to its predecessor HEVC on ultra high definition (UHD, a.k.a. 4K, 3840×2160) standard dynamic range (SDR) video content using formal subjective visual quality assessment testing. The purpose of the verification test was to confirm that the coding efficiency objective for the VVC standard has been met: achieving a substantial bit-rate reduction for the same level of *subjective* visual quality relative to the HEVC Main Profile. As anticipated in the test plan, in addition to using the HM reference software encoder for HEVC and the VTM reference software encoder for VVC, which used essentially the same rate-distortion optimization encoding techniques, another VVC encoder that uses alternative techniques for subjective quality optimization and faster encoding had also become available for study and was included in the test – namely the VVenC open-source VVC encoder. The VVenC encoder, although still a preliminary implementation produced only two months following the completion of the standard, was used to represent an example of practical encoding as may be found in product implementations and is reported to be more than 100 times faster than the VTM encoder. The compression performance of the HEVC reference software HM-16.22, the VVC reference software VTM-10.0, and the open-source VVC implementation VVenC-0.1.0, were compared for UHD SDR content using a random-access (RA) configuration suitable for streaming or broadcast applications.

The testing used the degradation category rating (DCR) test method (as in ITU-T P.910) with an 11-point impairment scale (as in Rec. ITU-R BT.500). The results of a visual assessment of VVC compared to HEVC by naïve test subjects are reported. The assessment included five test sequences encoded in a random-access configuration with a random-access interval of 1.07 seconds. The measured mean opinion score (MOS) figures indicate a significant improvement of VVC relative to HEVC for both VVC implementations, VTM-10.0 and VVenC-0.1.0, resulting in overall average bit-rate savings estimates of 43% and 49%, respectively.

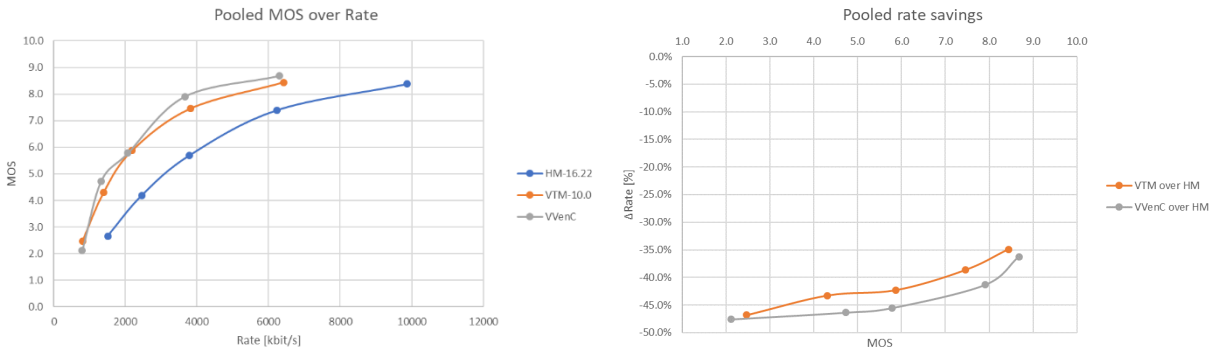


Figure 1: Pooled MOS over bit rate plot (left) and pooled bit-rate savings percentage over MOS for the five UHD SDR test sequences

1 Introduction

A major design goal for the development of the VVC standard was to achieve a substantial improvement in compression capability relative to its predecessor, the HEVC standard. This document is the first in a planned series of reports addressing a variety of test categories and embracing some of the available versatile tools provided by the VVC standard. It reports the results of a verification test to confirm that this goal was achieved and to estimate the magnitude of that achievement, following a test plan issued at the previous meeting [1].

A subjective evaluation was conducted at two test sites comparing the VVC Main 10 profile to the HEVC Main 10 profile for the UHD SDR test category with random-access configuration.

2 Verification test logistics

The UHD SDR subjective test was carried out at the following test sites:

- GBTech, Rome, IT
- RWTH Aachen University, Aachen, DE

The tests were conducted using the degradation category rating (DCR) test method [2] with an 11-grade impairment scale [3]. The verification test environment and testing methodology are described in Annex A. The arrangements for the two test sites are shown in Table 1.

Table 1: Test site information and setup

Test Site	**GBTech**	**RWTH Aachen University**
Display, size, connection (resolution setting)	LG 65” CX6LA, HDMI (3840×2160)	Sony 55” PVM X550, Quad-SDI (3840×2160)
Viewing distance	2 viewers at 1.5H	1 viewer at 1.5H
Viewing angle	60° (30° from screen center)	90° (at screen center)
Total number of viewers	16 (7 female, 9 male; ages 18-24) all screened for visual acuity and normal colour vision	24 (5 female, 19 male; ages 16-34) all screened for visual acuity and normal colour vision

3 Verification test sequences, encodings, and MOS results

3.1 Test sequences and encodings

In the test, the HEVC bitstreams were encoded using the HEVC reference software HM16.22 [4]. For VVC, two encoder implementations were used. One set of bitstreams was encoded using the VTM-9.0 reference software [5], a second set of bitstreams was encoded using the open source VVC implementation VVenC-0.1.0 [6][7][8]. For all, the configuration enables random access to the bitstream every 1.07 seconds, i.e. every 32 pictures for 30 Hz test sequences and every 64 pictures for 60 Hz test sequences.

For HEVC, the random-access configuration provided with the configuration file cfg/encoder_randomaccess_main10.cfg of HM-16.22 was used. For the VTM, the random-access configuration provided with the configuration file cfg/encoder_randomaccess_vtm_gop32.cfg of VTM-9.0 was employed. These selected HM and VTM configurations result in the application of very similar configurations and very similar searching and rate-distortion optimization techniques in the HEVC and VVC contexts (using fixed QP settings and greedy optimization techniques with Lagrange multiplier $D + \lambda \cdot R$ decision making), thus maximizing the ability to test the capability of the differing syntax and decoding process features of the tested HEVC and VVC profiles in a controlled manner.

It is also noted that the encoder decisions of VTM-10.0 have been asserted to be identical to those of VTM-9.0. The bit rate of the two VTM versions is very slightly different due to modifications of high-level signalling in VTM-10.0 which reflects the final version of VVC, but the reconstructed $Y'C_BC_R$ video sequences were asserted to be identical for VTM-10.0 and VTM-9.0 using this configuration [9]. Therefore, the VTM results are reported as VTM-10.0 results in this document. For the VVenc-0.1.0 software [6][7][8], the Medium configuration has been used. VVenC was operated with perceptual QP adaptation and without rate control. In this configuration, the VVenC-0.1.0 encoding speed is reportedly more than 100 times faster than the VTM [9]. For the purpose of the verification tests, the QP values were selected such that the VVenC bit rate was approximately the same or lower than the VTM bit rate.

The employed set of five test sequence candidates with UHD resolution and standard dynamic range as defined in the verification test plan document [1] is listed in Table 2.

Five bit-rate points for each test sequence were selected for the quality assessment of the test sequences. The bit-rate points were chosen such that the VTM/HM pair for a bit-rate point would represent approximately the same quality while at the same time allowing for approximate bit-rate matching of each HM bit-rate point with the next VTM bit-rate point. Thereby both an assessment of bit-rate savings at similar quality and an assessment of quality improvement at similar bit rates are enabled. The selected QPs corresponding to these bit-rate points are listed in Table 3.

Table 2: UHD SDR test sequences

No.	Test sequence	Resolution	fps	Frames	md5
01	DrivingPOV3	3840×2160	60	0:599	e81b65724c4235128b2749ccb3b0fb4a
02	Marathon2	3840×2160	30	0:299	c065dfb87be3b2e2ab0ce35094fd4eb4
03	MountainBay2	3840×2160	30	0:299	f27b6b70244fb083baac546958fcf696
04	NeptuneFountain3	3840×2160	60	0:599	88fd87ea57df4a36200946025e8618aa
05	TallBuildings2	3840×2160	30	0:299	9a0a3f261d004fa86754751c82fb8b47

Table 3: QP settings for HM, VTM, and VVenC for the UHD SDR test sequences

No.	Test sequence	HM-16.22 QP	VTM-9.0	VVenC-0.1.0
01	DrivingPOV3	30, 33, 36, 39, 43	30, 34, 39, 42, 46	29, 33, 38, 42, 46
02	Marathon2	28, 32, 36, 39, 43	30, 34, 39, 42, 46	29, 33, 38, 41, 45
03	MountainBay2	27, 30, 33, 36, 39	30, 33, 36, 39, 42	30, 33, 36, 39, 43
04	NeptuneFountain3	30, 32, 35, 37, 39	33, 35, 37, 39, 42	32, 34, 36, 38, 43
05	TallBuildings2	28, 31, 35, 39, 43	30, 34, 38, 42, 46	28, 32, 36, 40, 44

The test sequences were provided in the Rec. ITU-R BT.709 colour space [12][13].

The test sequences were evaluated using the 11-grade scale as specified in Rec. ITU-R BT.500-14, shown in Figure 2 below.

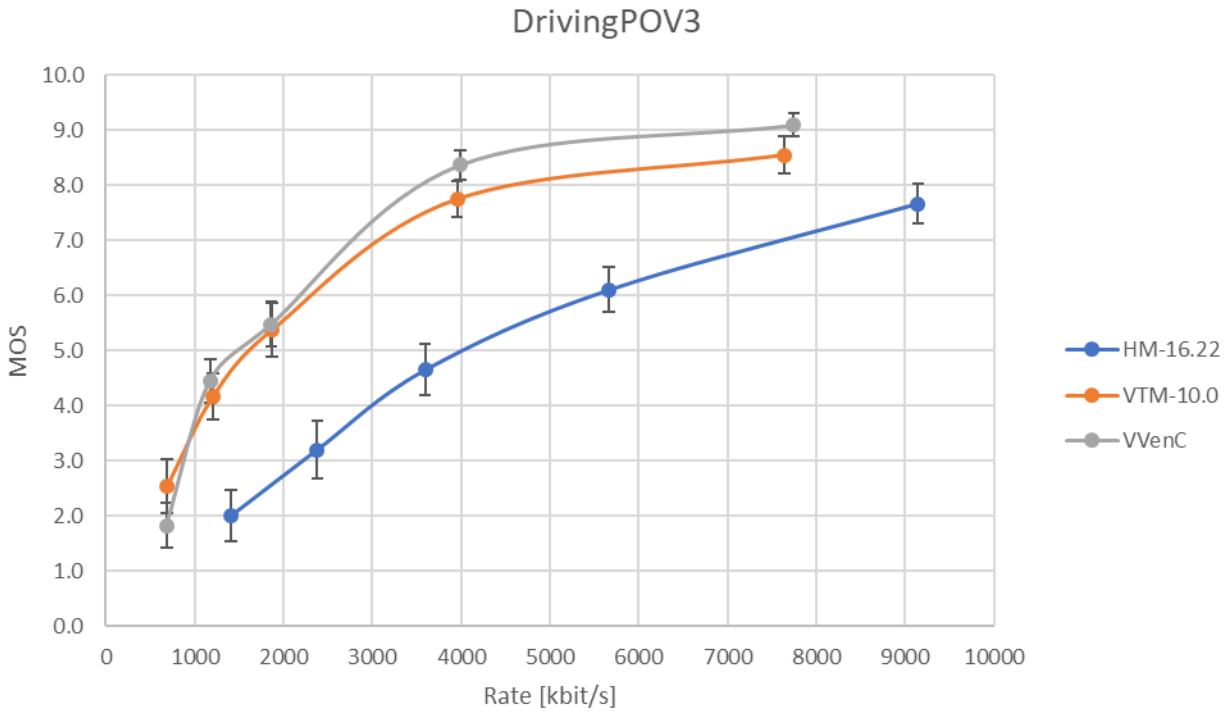
Score	Impairment item	
10	Imperceptible	
9	Slightly perceptible	somewhere
8		everywhere
7	Perceptible	somewhere
6		everywhere
5	Clearly perceptible	somewhere
4		everywhere
3	Annoying	somewhere
2		everywhere
1	Severely annoying	somewhere
0		everywhere

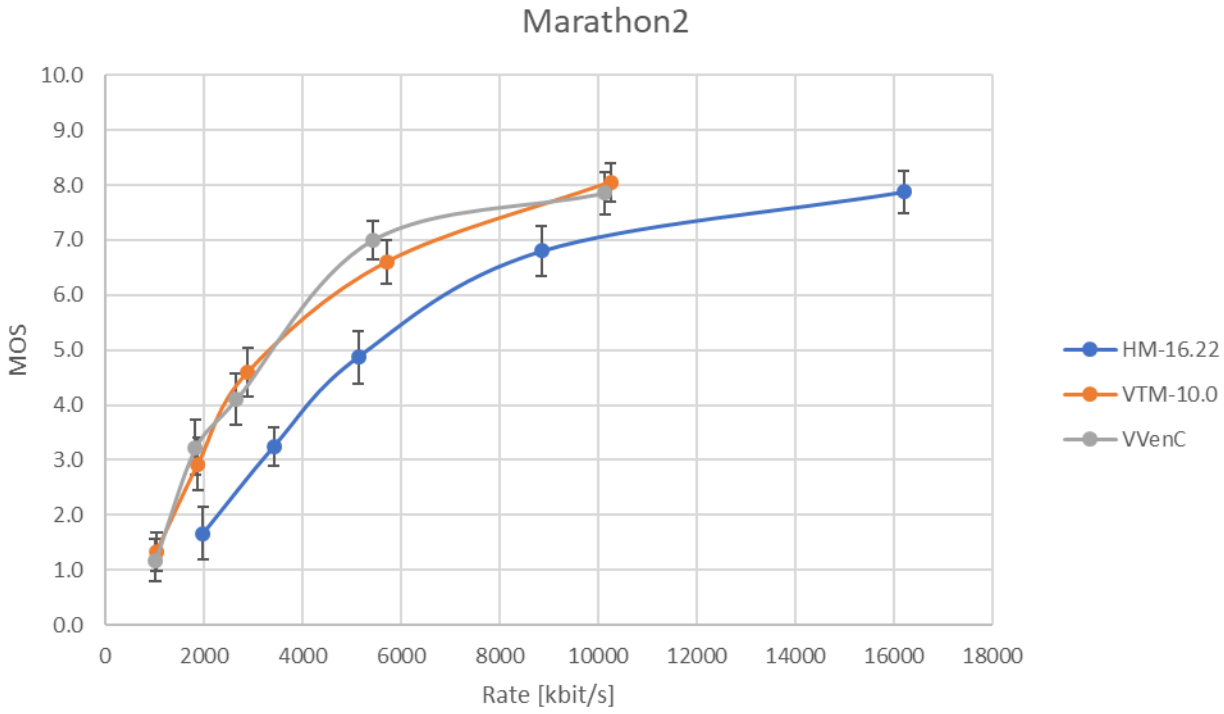
Figure 2: Meaning of the 11 grades numerical scale as specified in Rec. ITU-R BT.500-14 Table 2-4 [2]

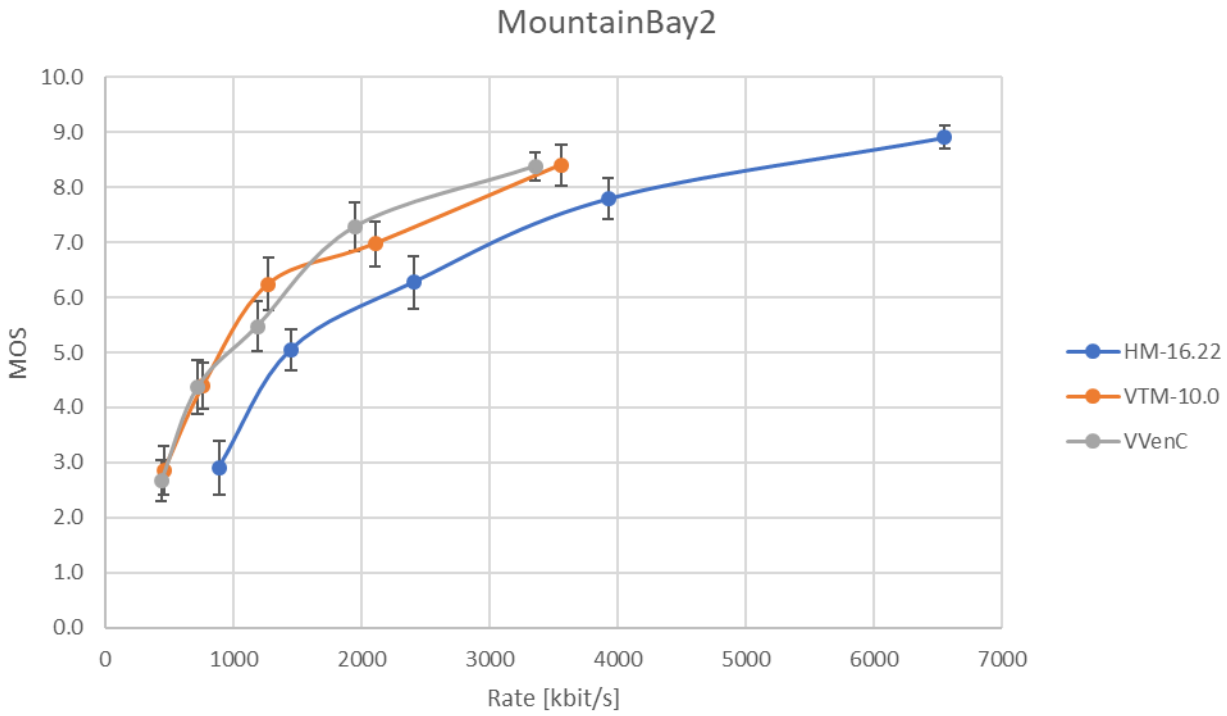
3.2 MOS plots

The results of the visual assessment at the two test sites were evaluated by the verification test coordinators and were found to be consistent in terms of quality range and quality progression over the bit-rate points as voted by the test subjects. Therefore, the votes of both test sites were merged into a joint evaluation providing the results provided below. The measured MOS values of the reconstructed video on the 11-grade scale are plotted over the bit rate of the corresponding bitstream. The ±0.95% confidence intervals for the MOS values are indicated.

The pooled plots in the executive summary have been generated by computing the geometric mean bit rate for each bit-rate point and the arithmetic mean of the corresponding MOS values.


DrivingPOV3
MOS
Rate [kbit/s]
HM-16.22
VTM-10.0
VVenC


Marathon2
MOS
Rate [kbit/s]
HM-16.22
VTM-10.0
VVenC


MountainBay2
MOS
Rate [kbit/s]
HM-16.22
VTM-10.0
VVenC

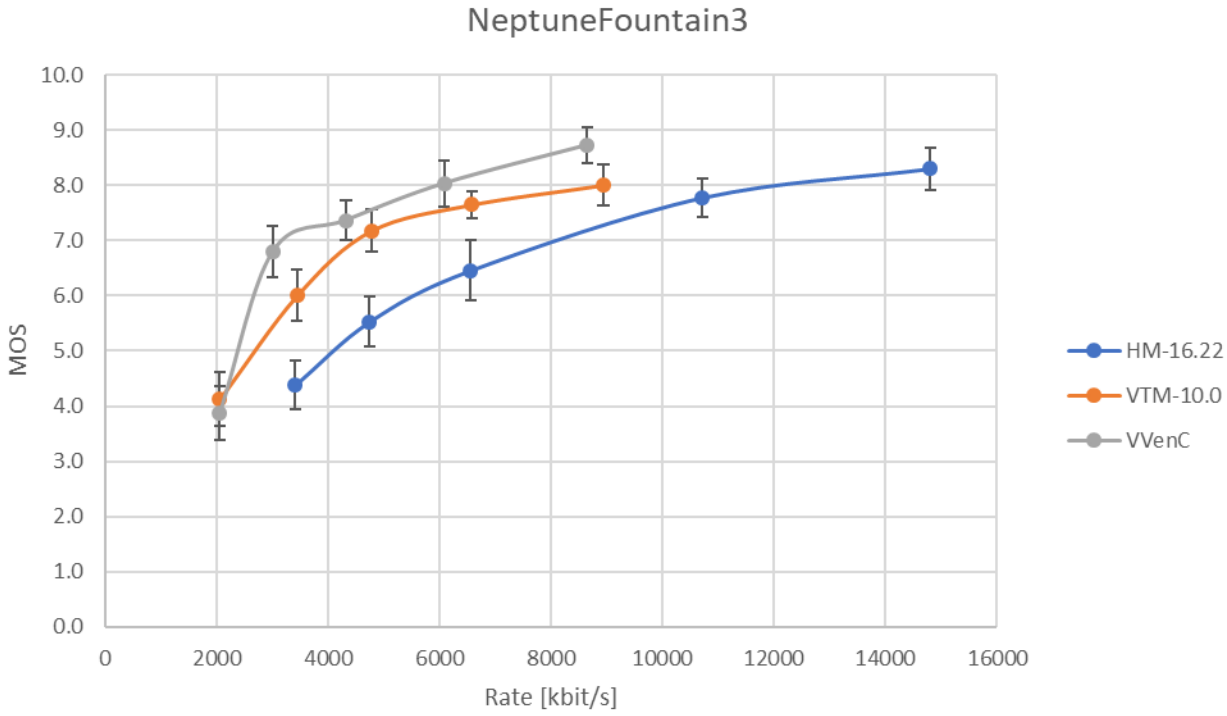

NeptuneFountain3
MOS
Rate [kbit/s]
HM-16.22
VTM-10.0
VVenC

Figure 3: Collection of the MOS-over-rate plots for HM, VVC, and VVenC for the UHD SDR test sequences

The MOS-over-rate plots provided in Figure 3 consistently indicate a significant subjective quality improvement of VVC over HEVC. For some sequences, a visual quality close to transparent (MOS higher than 8.5) is achieved already for the 4th bit-rate point of VTM and/or VVenC. The plots further reveal that for some test sequences and some bit-rate points, the intended quality matching of VTM and HM is not achieved in the subjective evaluation with naïve test subjects. This specifically holds for the DrivingPOV sequence, where the visual quality of the VVC encoders was consistently rated higher than the quality of the corresponding HM bit-rate points.

4 Analysis of the MOS over bit rate plots

Bjøntegaard delta rate (BD-rate) measurements [14][15][16] computed from the MOS results for both the VTM-9.0 and VVenC relative to HM-16.22 are reported in Table 4. The numbers have been calculated based using the 5-point BD method provided in JVET-T0041 [17].

The bit rate and MOS differences for all bit-rate points are collected in Table 5. The bit-rate savings is computed as the difference between the VVC bit-rate point and the corresponding HEVC bit-rate point relative to the HEVC bit rate. The MOS difference is reported as a number if the value is larger than the maximum of the VVC and the HEVC confidence intervals. Otherwise, "< CI" is indicated. The results are reported relative to the HM for both the VTM and VVenC encoders.

Table 4: Bjøntegaard delta rate relative to HM-16.22 based on bit rate and MOS

BD-Rate	**VTM**	**VVenC**
DrivingPOV3	−61%	−63%
Marathon2	−37%	−42%
MountainBay2	−37%	−39%
NeptuneFountain3	−38%	−52%
TallBuildings2	−41%	−51%
Overall	−43%	−49%

Table 5: Bit-rate savings and MOS deltas for the bit-rate points

VTM / HM		Rate Diff.	ΔMOS
DrivingPOV3	R1	−51.1%	0.5
DrivingPOV3	R2	−49.4%	1.0
DrivingPOV3	R3	−48.0%	0.7
DrivingPOV3	R4	−30.1%	1.7
DrivingPOV3	R5	−16.5%	0.9
Marathon2	R1	−47.7%	< CI
Marathon2	R2	−45.0%	< CI
Marathon2	R3	−43.6%	< CI
Marathon2	R4	−35.5%	< CI
Marathon2	R5	−36.7%	< CI
MountainBay2	R1	−48.8%	< CI
MountainBay2	R2	−47.5%	−0.6
MountainBay2	R3	−47.1%	< CI
MountainBay2	R4	−46.4%	−0.8
MountainBay2	R5	−45.7%	−0.5
NeptuneFountain3	R1	−39.8%	< CI
NeptuneFountain3	R2	−27.4%	0.5
NeptuneFountain3	R3	−27.1%	0.7
NeptuneFountain3	R4	−38.6%	< CI
NeptuneFountain3	R5	−39.6%	< CI
TallBuildings2	R1	−46.1%	−0.9
TallBuildings2	R2	−44.7%	< CI
TallBuildings2	R3	−43.3%	< CI
TallBuildings2	R4	−41.6%	< CI
TallBuildings2	R5	−32.6%	< CI

VCenC / HM		Rate Diff.	ΔMOS
DrivingPOV3	R1	**−51.2%**	< CI
DrivingPOV3	R2	**−50.6%**	1.3
DrivingPOV3	R3	−48.3%	0.8
DrivingPOV3	R4	−29.5%	2.3
DrivingPOV3	R5	−15.3%	1.4
Marathon2	R1	−48.7%	−0.5
Marathon2	R2	−46.4%	< CI
Marathon2	R3	−48.3%	−0.8
Marathon2	R4	−38.7%	< CI
Marathon2	R5	−37.4%	< CI
MountainBay2	R1	**−50.9%**	< CI
MountainBay2	R2	**−50.3%**	−0.7
MountainBay2	R3	**−50.7%**	−0.8
MountainBay2	R4	**−50.4%**	−0.5
MountainBay2	R5	−48.8%	−0.5
NeptuneFountain3	R1	−39.8%	−0.5
NeptuneFountain3	R2	−36.7%	1.3
NeptuneFountain3	R3	−34.1%	0.9
NeptuneFountain3	R4	−43.3%	< CI
NeptuneFountain3	R5	−41.6%	0.4
TallBuildings2	R1	−46.6%	−1.4
TallBuildings2	R2	−46.9%	0.8
TallBuildings2	R3	−44.8%	< CI
TallBuildings2	R4	−43.0%	< CI
TallBuildings2	R5	−33.6%	< CI

The figures reported in Tables 4 and 5 indicate significant compression performance improvements for VVC compared to its predecessor HEVC. It is observed that the VVC encoder implementation provided by VVenC provides more gain over the HM than the VTM for the vast majority of bit-rate points.

5 Conclusions

The results of a visual assessment of two VVC encoders compared to an HEVC Main 10 profile reference software encoder by naïve test subjects are reported. The assessment included five UHD SDR test sequences encoded in random-access configuration with a random-access interval of 1.07 seconds. The measured MOS figures indicate a significant improvement of VVC over HEVC for both VVC implementations, VTM-10.0 and VVenC-0.1.0, resulting in an overall average bit-rate savings of 43% and 49%, respectively.

6 References

[1] M. Wien, V. Baroncini, A. Segall, and Y. Ye, "VVC verification test plan (Draft 3)," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 output document JVET-S2009, 19th JVET meeting, by teleconference, June 2020.

[2] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[3] Recommendation ITU-R BT.500-14 (2019), *Methodologies for the subjective assessment of the quality of television images*.

[4] HEVC Test Model, https://vcgit.hhi.fraunhofer.de/jct-vc/HM/-/tree/HM-16.22, online, accessed Oct. 2020.

[5] VVC Test Model, https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-9.0, online, accessed Oct. 2020.

[6] J. Brandenburg, A. Wieckowski, T. Hinz, A. Henkel, V. George, I. Zupancic, C. Stoffers, B. Bross, H. Schwarz, and D. Marpe, "Towards Fast and Efficient VVC Encoding," IEEE 22nd Workshop on Multimedia Signal Processing (MMSP 2020), Tampere, Finland, 2020.

[7] A. Wieckowski, J. Brandenburg, C. Bartnik, V. George, J. Güther, G. Hege, C. Helmrich, A. Henkel, T. Hinz, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, H. Schwarz, and D. Marpe, "Open optimized VVC encoder (VVenC) and decoder (VVdeC) implementations," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 input document JVET-T0099, 20th JVET meeting, by teleconference, Oct. 2020.

[8] Fraunhofer HHI, "Fraunhofer Versatile Video Encoder (VVenC)," version 0.1.0 (initial release), https://github.com/fraunhoferhhi/vvenc/tags, online, accessed Oct. 2020.

[9] C. Helmrich, B. Bross, J. Pfaff, H. Schwarz, D. Marpe, and T. Wiegand, "Information on and analysis of the VVC encoders in the UHD SDR verification test," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 input document JVET-T0103, 20th JVET meeting, by teleconference, Oct. 2020.

[10] Recommendation ITU-R BT.2100-2 (2018), *Image parameter values for high dynamic range television for use in production and international programme exchange*.

[11] SMPTE ST 2084, *High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*, 2014.

[12] SMPTE ST 2036-1, *Ultra High Definition Television – Image Parameter Values for Program Production*, 2014.

[13] ETSI TS 101 154, *Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcast and Broadband Applications*, 2019.

[14] ITU-T Tech. Paper HSTP-VID-WPOM and ISO/IEC TR 23008-8 (*Eds.*: K. Andersson, F. Bossen, J.-R. Ohm, A. Segall, R. Sjöberg, J. Ström, G. J. Sullivan, A. Tourapis), Working practices using objective metrics for evaluation of video coding efficiency experiments, July 2020.

[15] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," ITU-T SG 16 Q.6 input document VCEG-M33, 13th VCEG meeting, Austin, Texas, USA, Apr. 2001.

[16] G. Bjøntegaard, "Improvements of the BD-PSNR model," ITU-T SG16 Q.6 document VCEG-AI11, 35th VCEG meeting, Berlin, Germany, July 2008.

[17] S. Liu, A. Segall, E. Alshina, J. Boyce, M. Wien, D. Grois, "Methodology and reporting template for neural network coding tool testing," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 input document JVET-T0041, 20th JVET meeting, by teleconference, Oct. 2020.

Acknowledgements

The verification test coordinators wish to thank the organizations who contributed to and supported the VVC verification tests, particularly including Alibaba Inc., Bytedance Inc., Fraunhofer HHI, GBTech, Huawei, MediaTek, RWTH Aachen University, Sharp Labs of America, and Tencent.

Annex A

The same evaluation method as for the HEVC verification tests is adopted for the VVC verification tests. The following description is based on JCTVC-Q1011 [A4] with minor adaptations.

A.1 Test method

The test method adopted for this evaluation is degradation category rating (DCR) [A1].

A.1.1 Degradation Category Rating (DCR)

This test method is commonly adopted when the material to be evaluated shows a range of visual quality that well distributes across all quality scales. All the video material used for these tests consist of video clips of 10 seconds duration.

This method has been used under the schema of evaluation of the quality; for this reason, a quality rating scale made of 11 levels was adopted, ranging from "0" (lowest quality) to "10" (highest quality), see also Figure 2.

The structure of the basic test cell (BTC) of the DCR method was made by two consecutive presentations of the video clip under test; at first the original version of the video clip is displayed, immediately afterwards the coded version of the video clip is presented; then a message displays for 5 seconds asking the viewers to vote. The presentation of the video clips is preceded by a mid-grey screen displaying “Source” for the original and “Test” for the coded version of the sequence under test for one second.

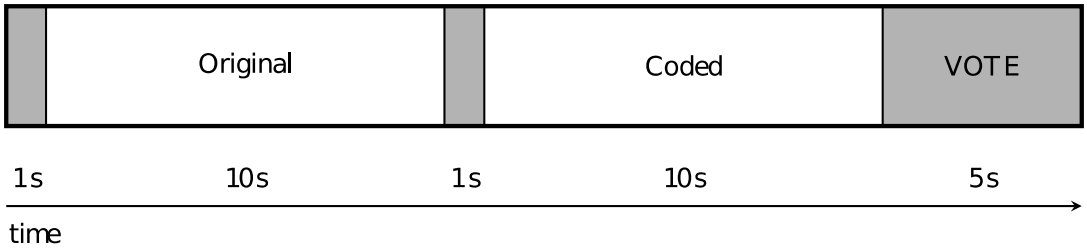


Figure 4 – DCR BTC

A.2 How to express the visual quality opinion with DCR

The viewers were asked to express their vote putting a mark on a scoring sheet.

The scoring sheet for a DCR test is made of a section for each BTC; each section has a box wherein which the viewer shall write the score ranging from 0 to 10. By writing a score of “10”, the subject will express an opinion of “best” quality, while by writing a score of “0” the subject will express an opinion of “worst” quality, as shown in Figure 2.

The vote has to be written when the message "Vote N" appears on the screen. The number "N" is a numerical progressive indication on the screen aiming to help the viewing subjects to use the appropriate box of the scoring sheet.

A.4 Training and stabilization phase

The outcome of a test is highly dependent on a proper training of the test subjects.

For this purpose, each subject has to be trained by means of a short practice (training) session demonstrating the range of qualities to be expected in the test.

The stabilization phase uses the test material of a test session; three BTCs, containing one sample of best quality, one of the worst qualities and one of medium quality, are duplicated at the beginning of the test session. By this way, the test subjects have an immediate impression of the quality range they are expected to evaluate during that session.

The scores of the stabilization phase are discarded.

A.5 The laboratory setup

The laboratories for subjective assessments were arranged according to [A1], except for the selection of the display and the video play-out server. Play-out of the UHD video clips was done at the native resolution.

The PCs used to play the video sequence supported the display of 10 bit UHD at 30 and 60 frames per second, without any limitation, or without introducing any additional temporal or visual degradation. At GBTech, the connection between the PC and the display was provided by a 10 bit-capable HDMI connection. At RWTH Aachen University, the display was connected via quad-link SDI.

A.5.2 Viewing environment

The viewing distance was 1.5H, where H is equal to the height of the active part of the screen, depending on the size of the active part of the screen and its native resolution.

The test laboratories were protected from external visual or audio pollution. Internal general light was low (just enough to allow the viewing subjects to fill out the scoring sheets) and a uniform light was placed behind the monitor, in a way no direct light hits the viewing subjects seated in front of the screen; the light behind the monitor must be dimmed to an intensity as specified in Table 4 of Recommendation ITU-T P.911 ("Typical viewing and listening conditions as used in audio-visual quality assessment"). No other light source was admitted, and in particular any light source directed to the screen or creating reflections.

A.6 Overall test effort and subjects' involvement

Each viewing session did not run for more than 20 minutes and the same viewing subject did not participate to the test run for more than six hours in total. Young people were hired as test subjects, selecting them for an age from 16 to 30, mostly students of scientific faculties. Viewing subjects were compensated for their participation to the testing activities.

A.7 Statistical analysis and presentation of the results

The data collected from the score sheets, filled out by the viewing subjects, were stored in an Excel spread sheet. For each coding condition the Mean Opinion Score (MOS) and associated Confidence Interval (CI) values were computed in the spread-sheets.

The MOS and CI values are used to draw graphs. The graphs are drawn grouping the results for each video test sequence. No graph grouping results from different video sequences is considered.

From the "raw" data subject reliability should be calculated and the method used to assess subject reliability should be reported. Some criteria for subjective reliability are given in [A2] and [A3].

A.8 References

[A1] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[A2] *Pseudo Isochromatic Plates*, engraved and printed by *The Beck Engraving Co., Inc*., Philadelphia and New York, United States.

[A3] KIRK (R.E.): *Experimental Design – Procedures for the Behavioural Sciences*, 2nd Edition, *Brooks/Cole Publishing Co*., California, 1982.

[A4] Tan, T. K.; Mrak, M.; Baroncini, V.; and Ramzan, N., "Report on HEVC compression performance verification testing," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 output document JCTVC-Q1011, Apr. 2014.