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-T0103-v1

Title: **Information on and analysis of the VVC encoders in the SDR UHD verification test**

Status: Input document to JVET

Purpose: Information

Author(s) or Contact(s): Christian Helmrich, Benjamin Bross, Jonathan Pfaff, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand

Email: christian.helmrich@hhi.fraunhofer.de

Fraunhofer HHI
Einsteinufer 37
10587 Berlin, Germany

Source: Fraunhofer HHI

Abstract

This contribution provides further information on the configuration of the VVC reference encoder (VTM) and the alternative encoder (VVenC) evaluated in the VVC verification test for SDR UHD content, specified in JVET-S2009 (test plan, draft 3). Based on an analysis of the bitstreams and Bjøntegaard delta-rate values obtained using several video quality assessment (VQA) methods, the following conclusions can be drawn:

- JVET-S2009 specifies VTM version 9.0 for the SDR UHD verification test, but the decoded video files resulting from newer VTM 10.0 encoded bitstreams were found to be identical to those resulting from the VTM 9.0 encodings. Thus, the SDR UHD verification test was, effectively, conducted using the VTM 10.0 software release, which was the latest VTM version as of late September 2020.
- The results of the SDR UHD verification test, reported in JVET-T0097, indicate that VVenC 0.1.0 (VTM 10 decoder compliant) tends to outperform the VTM encoder in terms of visual quality. It was found that only some of the VQA methods used for Bjøntegaard delta-rate evaluation – namely, the XPSNR, SSIM, and MS-SSIM methods – succeed in modelling this tendency with satisfactory accuracy. The PSNR and VMAF metrics, in particular, fail to model the subjective judgments well.

1 Introduction

The recently finalized VVC verification test for standard dynamic range (SDR) ultra-high definition (UHD) content, specified in draft 3 of the verification test plan [1], included two encoder variants under test: the VVC reference encoder software (VTM) in version 9.0 [2], [3] and an alternative encoder optimized for a higher level of visual quality, Fraunhofer's Versatile Video Encoder (VVenC) in version 0.1.0 [4], [5]. Both encoders are publicly available under open-source licensing terms and fully VVC standard compliant (the VTM streams are VTM 9.0 decoder compatible and the VVenC streams are VTM 10.0 decoder compatible). They were tested at five rate points ranging from a *nearly transparent* high rate (1) to a *low quality* rate (5).

This document provides additional information on the configurations of VTM and VVenC in the SDR UHD verification test and reports on an analysis performed to answer the following questions regarding the test:

- Is there any difference in perceptual quality between VTM 9 and 10 decoder compliant bitstreams?
- Which objective metrics can model the outcome of the subjective evaluation, reported in [6], best?

2 Difference Between VTM 9 and 10 Decoder Compliant Bitstreams

As noted in the previous section, the VTM encoded sequences in the UHD SDR verification test are VVC draft 9 compliant, while the VVenC encoded sequences are VVC draft 10 (syntax complete) compliant [2]. A comparative analysis of VTM 9 and VTM 10 Random Access (RA) bitstreams, generated according to draft 3 of the verification test plan [1] with a group of picture (GOP) size of 32 frames, was conducted. The goal of this analysis was to assess whether the VTM version affects the bitrate or video reconstruction data.

It was found that, although the VTM 10 bitstreams are between 0.01 % (rate point 1) and 0.08 % (rate point 5) smaller than their VTM 9 counterparts due to improvements in the high-level coding syntax [7], the decoded videos are, for all five video sequences included in the SDR UHD verification test, completely identical.

In other words, there is **no difference** between VTM 9 and VTM 10 encodings with regard to the decoded YUV files presented to the participants of the SDR UHD verification test during the viewing sessions. The only cause for differences in perceptual quality is, therefore, the choice of the encoder, i. e., VTM or VVenC.

Regarding differences in visual quality between VTM 10 (or 9) encodings and VVenC 0.1.0 encodings, [6] indicates that the latter tends to outperform the former encoder. Specifically, on roughly **90 %** of the bitrate range under test, the MOS/bit-rate curve for VVenC lies above the MOS/bit-rate curve for VTM, and on 4 of the 25 pairs of sequence-and-rate conditions evaluated (two rate points each on *DrivingPOV3* and *NeptuneFountain3*), VVenC provides **better** visual quality with statistical significance (95 % confidence level).

For the sake of documentation (this information is not included in [6], [8]), a summary of the configuration of the VVenC encoder, as used during the generation of the alternative bitstreams for inclusion in the SDR UHD verification test, follows. The VVenC encodings follow the requirements specified in the test plan [1] and include optimizations for subjective video quality and encoder-only efficiency improvements, namely,

- signal and in-GOP frame index dependent perceptual QP adaptation (QPA) based on the published JVET documents H0047 [9], K0206 [10], and M0091 [11], QPA related ALF encoder changes [12], and more recent XPSNR related enhancements to the spatiotemporal QPA algorithm, see section 3,
- encoder-only deblocking optimization (EDO) accounting for the effect of deblocking on the reconstructed-video signal distortion within the encoding parameter search-loop of each coded frame [13],
- encoder-only GOP based motion compensated temporal filtering (MCTF) as a denoising pre-filter to improve the motion prediction performance and thus provide objective and subjective gains [14].

The perceptual QPA algorithm, which adapts the QPs on the CTU level for UHD input, was activated using the command-line option *--PerceptQPA=2*. For the 60-Hz video files, option *--PerceptQPATempFiltIPic=1* was additionally specified. As speed-vs-quality tradeoff, the options *--preset medium --threads 6* were used, which results in multithreaded operation and a VVenC runtime that, with UHD SDR input, is approximately **110 times faster** than the VTM encoder runtime [8]. Preset *medium* automatically enables EDO and MCTF.

Note that in VTM, there is at present no predefined MCTF configuration for a GOP size of 32 frames. Thus, such a predefined configuration, activated on the first three temporally hierarchical layers of a size-32 GOP, was introduced in VVenC (for a size-16 GOP, an activation on the first two temporal layers is used in both VTM and VVenC). Details can be found in file *source/Lib/vvenc/EncCfg.cpp* of the VVenC distribution [5].

3 Performance of VQA Methods in Modelling Verification Test Result

The previous section noted that the participants of the SDR UHD verification test tended to prefer the visual quality of the VVenC encoded videos over that of the VTM encoded videos. The mean opinion score (MOS) data indicate that the VTM encoder can save roughly **43 % bitrate** over HEVC at equal subjective quality while the VVenC encoder can save **almost 50 % bitrate** at the same quality (MOS Bjøntegaard delta-rate, see [6] and below). One popular question, particularly in the context of quality-of-service (QoS) or quality-of-experience (QoE) scenarios, is whether there exist any video quality assessment (VQA) algorithms, i. e., *objective metrics*, which are able to reflect such tendencies in the MOS data, i. e., the *subjective judgments*. If any exist, it is also interesting to assess *how well* these VQA algorithms model the subjective MOS data.

Within the Joint Video Experts Team, the approach adopted to quantify bitrate savings of some technology *A* relative to another technology *B*, while maintaining an equal level of distortion, is to calculate Bjøntegaard delta-rate (BDR) values [15]. To measure the coding distortion, the peak signal-to-noise ratio (PSNR) has been used traditionally, but other, perceptually motivated VQA methods may be used instead. Specifically,

- the structural similarity (SSIM) and multiscale structural similarity (MS-SSIM) measures [15],
- the video multimethod assessment fusion (VMAF) method combining several VQA models [16],
- and the extended perceptually weighted peak signal-to-noise ratio (XPSNR) method [17], [18], [19]

are widely used and/or readily available in software. The XPSNR algorithm represents a generalization of the PSNR metric, with a block-wise measure of *sum of squared errors* (SSE) distortion locally *weighted* by

$$w_k = \sqrt{\frac{a_{\text{pic}}}{a_k}},\ \ a_{\text{pic}} = 2^{2BD-9} \cdot \sqrt{\frac{3840 \cdot 2160}{W \cdot H}},\ \ a_k = \max\left(2^{2BD-12}, \left(\frac{1}{4N^2}\sum_{[x,y]\in B_k} |h_{\text{s}}[x,y]| + 2|h_{\text{t}}[x,y]|\right)^2\right),$$

where a_{pic} is a video bit depth (*BD*), width (*W*) and height (*H*) dependent constant and the block-wise visual activity measure a_k is defined as in [18] for each *N*×*N* sized block at index *k*. For an entire video sequence,

$$XPSNR = 20 \cdot \log_{10}\left(\frac{F \cdot \sqrt{W \cdot H} \cdot (2^{BD}-1)}{\sum_{i=1}^{F} \sqrt{\sum_k WSSE_k}}\right),\ \ WSSE_k = w_k \cdot \sum_{[x,y]\in B_k} (s_{\text{dec}}[x,y] - s[x,y])^2,$$

where *F* equals the number of frames in the video, $WSSE_k$ is the picture and block-wise weighted SSE, and s_{dec} and s are the decoded and original picture data, respectively. The picture/block-wise visual sensitivity weight w_k is used in VVenC as a simple psychovisual model to control the perceptual bit allocation (QPA) [18]. For HD (1920×1080) and UHD (3840×2160) input, this reuse of w_k is straightforward since the block size *N* was chosen to perfectly align with VVC's CTU size: *N* = 64 (quarter of CTU) for HD, 128 for UHD.

Per-sequence and overall BDR values for all five of the abovementioned objective measures were computed according to [15] on the SDR UHD verification test data, with HEVC encodings made using HM 16.22 [20] as the *reference* conditions and VVC encodings generated using VTM and VVenC as the *test* conditions, as noted earlier. Due to the use of five rate points per video sequence, the five-point BDR function defined in the JVET-T0041 Excel reporting template [21] was adopted for this purpose (macro *bdRateExtend*). PSNR averages across the color components were obtained using [6, 1, 1] / 8 weighting [15], and for XPSNR, the minimum of the component-wise output values (which in all cases equaled the value for luma) was selected. The SSIM, MS-SSIM, and VMAF (*4K* and *no enhancement gain* model) values were calculated using [16].

For VTM, the sequence-wise bitrates obtained using version 10.0 were used instead of those for version 9.0 since, as noted in section 2, it was found that the former were slightly smaller (the VTM 10 encodings result in a 0.02 % higher BDR gain over HM than the VTM 9 encodings on average for the UHD SDR test). SEI messages, including those for checksums of the reconstructed picture data, were disabled in all bitstreams.

The results of the "objective BDR" evaluation, listed in **Tables 1** and **2** and in further detail in the Excel file accompanying this document, reveal that the SSIM, MS-SSIM, and XPSNR measures succeed in modelling the higher BDR gain of the VVenC relative to the VTM encoder, as observed in the "subjective BDR" data. Despite being about 10 % lower than the MOS based BDR "ground truth" values, these VQA values predict the overall tendencies in the MOS data, i. e., the overall subjective codec ranking, with acceptable accuracy.

The PSNR and VMAF metrics, however, fail to reflect these tendencies, especially on the *NeptuneFountain* sequence. Regarding PSNR, which is not a sophisticated perceptually motivated distortion model, this issue is well known and can, therefore, be expected. The suboptimal performance of the VMAF metric, however, is somewhat surprising and may indicate a shortcoming in VMAF's psychovisual model. The issue appears to be most prominent on content with strongly irregular motion and/or low contrast, such as the shady water regions on *NeptuneFountain* (where VVenC's perceptual bit allocation method applies strong QPA). Hence, a possible reason for VMAF's sub-average performance may lie in an insufficient modelling of the temporal contrast sensitivity aspects of human vision. At present, the use of the (MS-)SSIM or XPSNR methods in addition to the PSNR metric for BDR evaluations may, therefore, be a "safer" option than the use of VMAF.

VTM 10.0	Over HEVC (HM 16.22)						
	PSNR YUV	XPSNR	VMAF-4K	VMAF-neg	MS-SSIM	SSIM	**MOS** [6]
DrivingPOV3	-42.75%	-44.77%	-51.07%	-49.34%	-44.64%	-45.46%	-60.87%
Marathon2	-32.93%	-35.81%	-36.21%	-35.40%	-34.69%	-31.63%	-37.18%
MountainBay2	-38.59%	-32.70%	-47.21%	-43.90%	-32.00%	-31.28%	-37.08%
NeptuneFountain3	-26.76%	-28.71%	-26.97%	-26.05%	-24.26%	-23.29%	-37.93%
TallBuildings2	-37.26%	-34.11%	-43.65%	-41.43%	-32.61%	-30.30%	-40.68%
Overall	-35.66%	-35.22%	-41.02%	-39.23%	-33.64%	-32.39%	**-42.75%**

Table 1: BDR results, relative to the HM encoder, of the VTM encoder in the SDR UHD verification test.

VVenC 0.1	Over HEVC (HM 16.22)						
	PSNR YUV	XPSNR	VMAF-4K	VMAF-neg	MS-SSIM	SSIM	**MOS** [6]
DrivingPOV3	-35.51%	-42.00%	-43.88%	-42.46%	-41.38%	-44.81%	-62.89%
Marathon2	-22.32%	-32.10%	-26.99%	-27.31%	-33.88%	-34.65%	-41.79%
MountainBay2	-38.82%	-37.41%	-40.91%	-39.60%	-33.07%	-34.46%	-38.81%
NeptuneFountain3	-16.20%	-35.43%	-11.42%	-13.42%	-24.85%	-31.56%	-51.91%
TallBuildings2	-34.06%	-35.98%	-37.24%	-36.50%	-38.01%	-40.92%	-50.98%
Overall	-29.38%	-36.58%	-32.09%	-31.86%	-34.24%	-37.28%	**-49.28%**

Table 2: BDR results, relative to the HM encoder, of the VVenC encoder in the SDR UHD verification test.

4 Summary and Conclusion

An additional analysis of the VTM (version 9.0 and 10.0) and VVenC (version 0.1.0) encodings evaluated in the VVC verification test for SDR UHD content [1], [6] allows for the following conclusions to be drawn:

1. The choice of VTM version 9 or 10 did not have any influence on the reconstructed video signals presented during the viewing sessions, only the bitstreams were found to be smaller for version 10.
2. Given *conclusion 1*, only the choice between the VVC encoding frameworks, i. e., VTM or VVenC, affected the visual coding quality among the encoders evaluated in the SDR UHD verification test. As indicated in [6], VVenC tends to outperform the VTM encoder in terms of visual coding quality.
3. Bjøntegaard delta-rates calculated using, as distortion measures, PSNR as well as four VQA models indicate that SSIM, MS-SSIM, and XPSNR are able to predict the subjective codec ranking reported in [6], as mentioned in *conclusion 2*, with acceptable accuracy. The PSNR and two VMAF measures were found not to model the subjective BDR tendencies between the evaluated encoders very well.

Recently, there have been discussions within the Joint Video Experts Team on the usage of VQA metrics in addition to PSNR during evaluation or standardization work [22]. The outcome of the investigation reported in this contribution can serve as an indicator that the SSIM, MS-SSIM, and XPSNR measures may represent candidates with potentially sufficient performance when it comes to objective prediction of subjective rankings between coding technologies. However, further studies should be conducted to verify these findings.

5 Patent rights declaration(s)

Fraunhofer HHI may have current or pending patent rights relating to the technology described in this contribution and, conditioned on reciprocity, is prepared to grant licenses under reasonable and non-discriminatory terms as necessary for implementation of the resulting ITU-T Recommendation | ISO/IEC International Standard (per box 2 of the ITU-T/ITU-R/ISO/IEC patent statement and licensing declaration form).

6 References

[1] M. Wien, V. Baroncini, A. Segall, Y. Ye, "VVC verification test plan (Draft 3)," Joint Video Experts Team, doc. JVET-S2009, version 2, teleconference, July 2020.

[2] B. Bross, J. Chen, S. Liu, Y.-K. Wang (editors), "Versatile Video Coding (Draft 10)," Joint Video Experts Team, doc. JVET-S2001, version 17, teleconference, July–Sep. 2019.

[3] JVET, "VVC VTM reference software," https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-10.0

[4] 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," in *Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, virt./online, Sep. 2020.

[5] Fraunhofer HHI, "Fraunhofer Versatile Video Encoder (VVenC)," version 0.1.0 (initial release), repository, Sep. 2020. https://github.com/fraunhoferhhi/vvenc/tags

[6] M. Wien, V. Baroncini, "Report on VVC compression performance verification testing in the SDR UHD Random Access Category," Joint Video Experts Team, doc. JVET-T0097, teleconf., Oct. 2020.

[7] F. Bossen, X. Li, K. Sühring, "JVET AHG report: Test model software development (AHG3)," Joint Video Experts Team, doc. JVET-T0003, teleconference, Oct. 2020.

[8] 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, D. Marpe, "Open optimized VVC encoder (VVenC) and decoder (VVdeC) implementations," Joint Video Experts Team, doc. JVET-T0099, teleconference, Oct. 2020.

[9] S. Bosse, C. Helmrich, H. Schwarz, D. Marpe, T. Wiegand, "Perceptually optimized QP adaptation and associated distortion measure," Joint Video Experts Team, doc. JVET-H0047, Macao, Oct. 2017.

[10] C. Helmrich, H. Schwarz, D. Marpe, T. Wiegand, "AHG10: Improved perceptually optimized QP adaptation and associated distortion measure," Joint Video Experts Team, doc. JVET-K0206, July 2018.

[11] C. Helmrich, H. Schwarz, D. Marpe, T. Wiegand, "AHG10: Clean-up and finalization of perceptually optimized QP adaptation method in VTM," Joint Video Experts Team, doc. JVET-M0091, Jan. 2019.

[12] C. Helmrich, B. Bross, J. Erfurt, "AHG10: Corrected operation of ALF encoding with perceptually optimized QP adaptation," Joint Video Experts Team, doc. JVET-L0181, version 2, Macao, Oct. 2018.

[13] N. Hu, V. Seregin, W.-J. Chien, M. Karczewicz, "Encoder optimization with deblocking filter," Joint Video Experts Team, doc. JVET-M0428, version 2, Marrakech, Jan. 2019.

[14] P. Wennersten, R. Sjöberg, J. Enhorn, "AHG10: Encoder-only GOP-based temporal filter," Joint Video Experts Team, doc. JVET-O0549, version 2, Gothenburg, July 2019.

[15] 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 (Draft 3)," Joint Video Experts Team, doc. JVET-S2016, teleconference, July 2020.

[16] Netflix, "VMAF – Video Multi-Method Assessment Fusion," version 1.3.15, repository, Sep. 2019. https://github.com/Netflix/vmaf/tags

[17] C. R. Helmrich, M. Siekmann, S. Becker, S. Bosse, D. Marpe, and T. Wiegand, "XPSNR: A Low-Complexity Extension of the Perceptually Weighted Peak Signal-to-Noise Ratio for High-Resolution Video Quality Assessment," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, virt./online, May 2020. https://ieeexplore.ieee.org/document/9054089/

[18] C. R. Helmrich, S. Bosse, H. Schwarz, D. Marpe, and T. Wiegand, "A Study of the Extended Perceptually Weighted Peak Signal-to-Noise Ratio (XPSNR) for Video Compression with Different Resolutions and Bit Depths," *ITU Journal: ICT Discoveries (special issue)*, vol. 3, no. 1, pp. 65–72, May 2020. http://handle.itu.int/11.1002/pub/8153d78b-en

[19] Fraunhofer HHI, "XPSNR Filter Plug-in for FFmpeg," version 1.0 (initial release), repository, Sep. 2020. https://github.com/fraunhoferhhi/xpsnr/tags

[20] JCT-VC, "HM HEVC reference software," https://vcgit.hhi.fraunhofer.de/jct-vc/HM/-/tags/HM-16.22

[21] 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, doc. JVET-T0041, version 3, teleconference, Aug. 2020.

[22] S. Liu, E. Alshina, J. Pfaff, M. Wien, P. Wu, Y. Ye, "Report of AHG11 Meeting on Neural Network-based Video Coding on 2020-07-30," Joint Video Experts Team, doc. JVET-T0042, teleconference, Aug. 2020.