|  |  |
| --- | --- |
| **Joint Video Experts Team (JVET)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  15th Meeting: Gothenburg, SE, 3–12 July 2019 | Document: JVET-O0451 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **AHG4:** Subjective comparison of VVC and HEVC | | |
| *Status:* | Input document | | |
| *Purpose:* | Information | | |
| *Author(s) or Contact(s):* | P. Philippe, W. Hamidouche, J. Fournier, J.Y. Aubié | Tel: Email: | +33 2 99 12 40 61 [pierrick.philippe@orange.com](mailto:pierrick.philippe@orange.com), [Wassim.Hamidouche@insa-rennes.fr](mailto:Wassim.Hamidouche@insa-rennes.fr) |
| *Source:* | Orange, INSA Rennes and b<>com | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

This document reports a subjective evaluation of VVC. A formal subjective test was organised to compare HEVC and the latest VVC version (VTM 5.0) on a variety of HD and UHD SDR video sequences for a wide range of bit rates.

Under these testing conditions, it is reported that the performance improvement of VVC over HEVC is in the range of 37% for HD and 40% in UHD.

# Introduction

VVC will be finalised in approximately one year and is likely reaching the Committee Draft at the July 2019 meeting. Considering this milestone, it is important to assess its level of performance, and corroborate the level of improvement offered by VVC relative to HEVC. A visual testing has been organised by b<>com, INSA of Rennes and Orange to evaluate the subjective improvements offered by the VTM-5.0 over the HM-16.20.

# Test design

## Test methodology

The SAMVIQ methodology [1], was selected for this subjective assessment. It is recommended for the subjective assessment of multimedia applications. It was successfully used in the past e.g. in [2] to assess the level of improvement offered by HEVC compared to AVC on HD content.

SAMVIQ is a dedicated method to determine the video quality by giving accurate quality scores, it uses a continuous quality scale (from 0 to 100) following the following quality levels.

|  |  |
| --- | --- |
| **Score** | **Quality** |
| 0-20 | Bad |
| 20-40 | Poor |
| 40-60 | Fair |
| 60-80 | Good |
| 80-100 | Excellent |

This methodology uses an explicit and a hidden reference, the hidden reference score is reported.

Additionally, the SAMVIQ methodology provides recommendations for the screening of users, the selection of test sequences and the reporting of the score.

## Rates and coding configuration

A set of bit rates was selected to reflect to reflect the current and expected practices in video coding with the upcoming VVC standard. This bit-rate range is also selected to cover a wide range of quality levels, for both critical and simpler video clips.

* for HD, the bit rate range goes from 500 kbps to 10 Mbps
* for UHD, it goes from 1 Mbps to 20 Mbps

Regular increments are provided in order to provide regular increase on the quality scale, giving a total of 9 rate points as follows:

|  |  |  |
| --- | --- | --- |
| **Rate** | **1080 (HD)** | **2160 (UHD)** |
| **R1** | 500 | 1000 |
| **R2** | 1000 | 2000 |
| **R3** | 1500 | 3000 |
| **R4** | 2000 | 4000 |
| **R5** | 3000 | 6000 |
| **R6** | 4000 | 8000 |
| **R7** | 6000 | 12000 |
| **R8** | 8000 | 16000 |
| **R9** | 10000 | 20000 |

The coding configuration was identical for HEVC and VVC. The coding parameters, as described in the Common Test Conditions, were used, setting Random Access point every 1s (within a multiple of 16 frames).

* For HEVC, the encoder is HM16.20, using the main10 profile in the Random-Access (RA) condition (encoder\_randomaccess\_main10.cfg).
* For VVC, the VTM-5.0 also uses the default RA profile (encoder\_randomaccess\_vtm.cfg).

The bitrate target was merely respected by tuning the base QP. The deviation with respect to the desired bitrates was within 8% and most of the time within 3%.

## Test Sequences

The target resolutions for this test are HD (i.e. 1920x1080) and UHD (3840x2160), in a Standard Dynamic Range (SDR).

An initial set of 13 sequences was initially selected with the following criteria:

* Available in 3840x2160 or higher resolutions
* Homogenous content and absence of scene cut
* Duration of 10s minimum, to allow a meaningful subjective evaluation.
* Variety of spatial and temporal content
* Variety of clip genre (sport, test scenes, surveillance, …)
* Balance of critical and non-critical items

Those 13 sequences were down-sampled using the SHVC down-sampling filters to obtain 1920x1080 clips for the HD test.

Those clips were encoded both with HEVC (HM16.20) and VVC (VTM-5.0) in HD and UHD format, in order to retain those representing a good balance of content variety and video coding artifacts.

After this initial selection, 7 clips were finally retained. Their characteristics is presented below, also a thumbnail for each video is displayed:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence** | **Frame Rate** | **Bit Depth** | **Source** |
| AerialCrowd2 | 30 | 10 | Huawei |
| CatRobot1 | 60 | 10 | b<>com |
| CrowdRun | 50 | 8 | SVT |
| DaylightRoad | 60 | 10 | Huawei |
| Drums2 | 50 | 10 | b<>com |
| HorseJumping | 50 | 10 | [4Ever2](http://www.4ever-2.com/) |
| Sedof | 60 | 8 | 4Ever |

AerialCrowd2



CatRobot1



CrowdRun



DaylightRoad



Drums2



HorseJumping



Sedof

# Encoding and objective metrics

After encoding with the proper configuration, in HD and UHD, with the HM16-20 and VTM-5.0 software, a per sequence and per codec bit rate range was selected in order to span the SAMVIQ quality scale.

In particular, bit rates offering quality level lower than the “Poor” grade were avoided, as this level quality is not desired for applications. This selection was done, during an informal viewing session, where some experts graded the quality for the nine rates selected.

The tables below list the selected operating points for each sequence and resolution. All the coded sequences were also assessed with some objective metrics, namely the Y-PSNR, the MS-SSIM and VMAF. These were computed using the VMAF tool provided in [3] after decoding each sequence individually.

To evaluate the relative compression efficiency of VVC over HEVC, the BD-Rate metrics are also computed based on the Y-PSNR and VMAF metrics. For VMAF, two models were used for HD and UHD content, respectively vmaf\_v0.6.1.pkl and vmaf\_4k\_v0.6.1.pkl.

## HD Coding results and objective metrics

The rate is expressed in kilobit per second (kbps), the Y-PSNR in dB and the VMAF metric is truncated to the nearest integer.















## UHD Coding results and objective metrics

The numbers are proved with the same units as above.















## Observations and summary of objective scores

As can be seen from the tables above, the overall QP range is 25-39 for HEVC and 26-43 for VVC, this is slightly above the usual JVET testing conditions.

Additionally, the average BD-Rate metric is reported for each resolution, this is presented below.



The level of objective improvement of VVC over HEVC is merely in line with the numbers reported with the sequences and QP from the VVC Common Test Conditions (in the range of 35%).

# Subjective Test set-up

## Methodology

The SAMVIQ methodology (Cf. instructions in Appendix) was used.

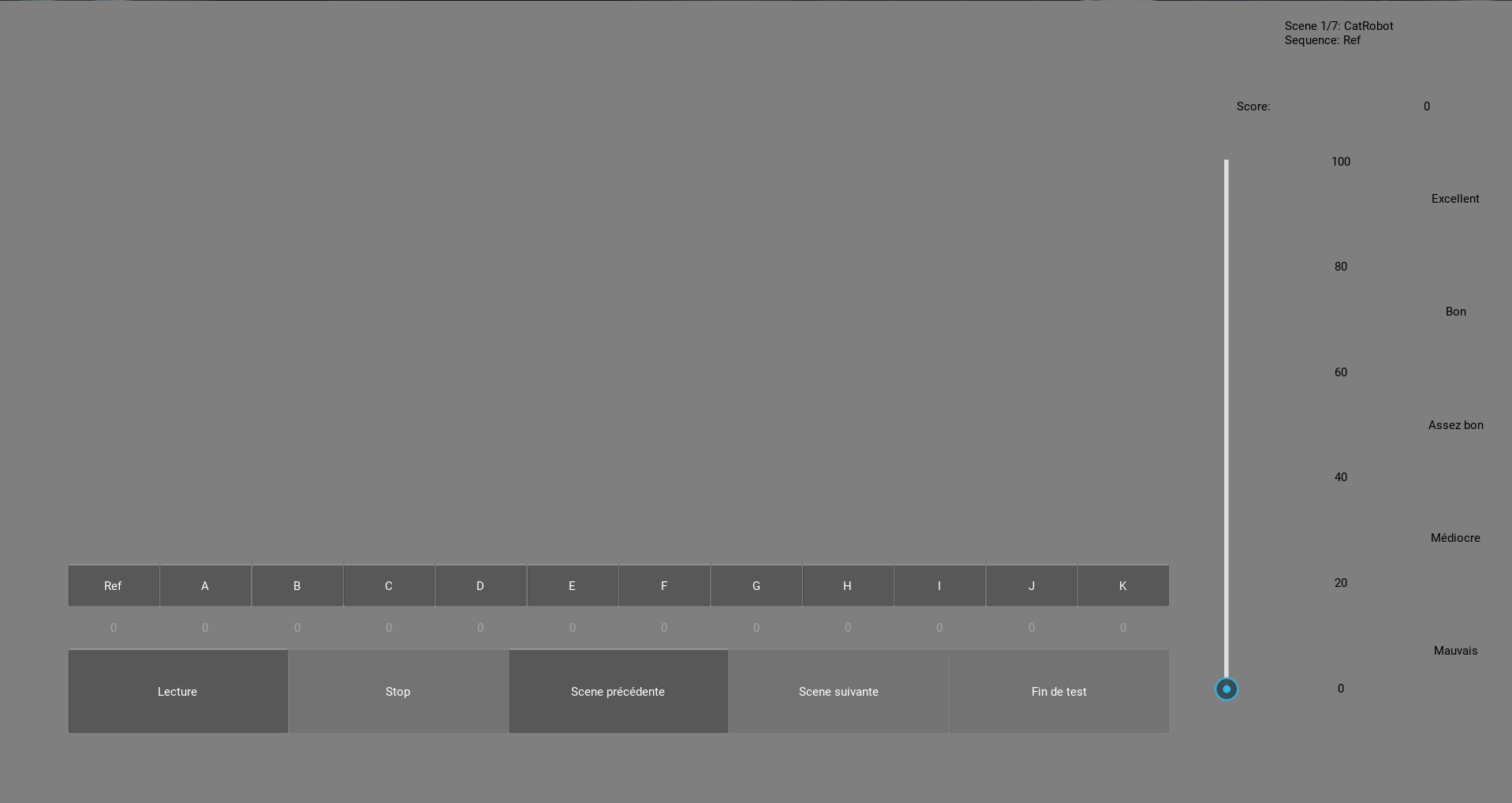
Twelve presentations were rated by the subjects:

* The explicit reference
* The hidden reference
* 5 HEVC coded items, at 5 rates
* 5 VVC coded items, at 5 rates

Consistent with the SAMVIQ methodology, the explicit reference was scored by the viewers, also the hidden reference was scored (no mention of the presence of a hidden reference was made to the subjects).

## Scoring and visual equipment

The figure below presents the interface of the scoring software.



The viewing conditions characteristics are displayed in the table below

|  |  |  |
| --- | --- | --- |
| Parameter | HD Test | UHD test |
| Monitor | Loewe Bild 7.55 | |
| Display Size | 55 inches | |
| Display Technology | OLED | |
| Peak luminance | 270 cd/m2 | |
| Viewing Distance | 3x the screen height | 1.5x the screen height |
| Room Size | ±12 square meters | |
| Background room illumination | 31 cd/m2 | |
| Video Display Card | DVS | |

The viewing setup is presented hereunder.



Mostly students were invited for the viewing session. The session was split in two sub-sessions, 3 scenes were rated then 4 scenes in order to avoid fatigue. The session started after a proper instruction and familiarisation with the scoring interface.

* The test is carried out scene by scene.
* For the current scene, it is possible to play and score any sequence in any order. Each sequence can be played and scored several times.
* From one scene to another, the sequence access is randomized and prevents the observers from attempting to vote in an identical way according to an established order.
* For a first viewing, the current sequence must be totally played before being scored; otherwise it would be possible to score and stop immediately.
* To test the next scene all sequences of the current scene must be scored.
* To finish the test all the sequences of all the scenes must be scored.

# Subjective test results

## HD test

For the HD test 23 naïve viewers were selected.

A post-screening was performed on the resulting subjects’ score. Following the SAMVIQ recommendation a post-screening based on the Pearson and Spearman rank correlations was operated.



Based on this screening, 5 subjects were discarded according to the correlation threshold (0.77 here).

A total of 18 observers were retained in the subsequent analysis.

The figures below present the MOS score obtained for each codec, the confidence intervals at 95% are also presented.















The following tables summarise the BD-Rate gains obtained by VVC relative to HEVC, it gives an indication of the compression gains, based on the subjective scores (MOS), on a per sequence basis.



Based on those numbers it seems that overall the gains are consistent with the objective measurement.

## UHD test

For the HD test 21 naïve viewers were selected.

A post-screening was performed on the resulting subjects’ score. Following the SAMVIQ recommendation, it is based on the Pearson and Spearman rank correlations.



Based on this screening, 2 subjects were removed from the statistical analysis. Their Spearman or Pearson correlation being lower than the threshold (0.728 here). Consequently, 19 viewers were retained in the subsequent statistical analysis.

The figures below present the MOS score obtained for each codec, the confidence intervals at 95% are also presented.















The following tables summarise the BD-Rate gains obtained by VVC relative to HEVC, it gives an indication of the compression gains, based on the subjective scores (MOS), on a per sequence basis.



# Conclusion

This document reports a subjective evaluation of VVC relative to HEVC. The SAMVIQ methodology was used.

The results show that VVC outperforms consistently HEVC, for realistic bit rates and quality levels, in the range of 40% on the subjective scale.

# References

1. SAMVIQ - Subjective assessment methodology for video quality, Recommendation ITU-R BT.1788
2. m29210, Subjective comparison of HEVC and AVC for HDTV content, Z. Agyo et al.
3. VMAF – Video Multi-Method Assessment Fusion, <https://github.com/Netflix/vmaf>

# Patent rights declaration(s)

**Orange, INSA de Rennes and b<>com 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).**

# Appendices

## Instruction for viewers

Welcome to INSA Rennes,

* You are about to take part in an evaluation of the quality of video sequences (video only, no sound). For each sequence, you have **to assess the overall video quality** for the entire duration. In addition, the visibility level of visual degradations can be used to assess the video quality.
* Seven various clips of about 10 seconds long have been selected (horse jumping, boat, …). Each of them has been treated with different processes indicated by the letters A, B, C, …. The reference clip (“REF” button) has not been processed.
* You may view each sequence in any order and repeat it as many time as you want (at least one time entire duration) using the “Play” button. After the visualization of each sequence, you can report your opinion moving the slider on the quality scale (numbered from 0 to 100) according to quality labels “Bad”, “Poor”, “Fair”, “Good”, “Excellent”.
* The scoring can be modified or refine at any time. You have to score the sequences of one clip before to assess the next clip pressing the “Next” button.
* At the end of the last sequence of the last clip, the “END” button becomes active. Press it to complete the test session.

***Thank for your participation.***