|  |  |
| --- | --- |
| **Joint Collaborative Team on Video Coding (JCT-VC)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  24th Meeting: Geneva, CH, 26 May – 1 June 2016 | Document: JCTVC-X1004 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Supplemental SHVC verification test report** | | |
| *Status:* | Output Document Approved by JCT-VC | | |
| *Purpose:* | Final test report | | |
| *Author(s) or Contact(s):* | Vittorio Baroncini  Walt Husak  Adarsh K. Ramasubramonian  Joel Sole  Gary J. Sullivan  Ye-Kui Wang  Yan Ye | Email: | [baroncini@gmx.com](mailto:baroncini@gmx.com)  [wjh@dolby.com](mailto:wjh@dolby.com)  [aramasub@qti.qualcomm.com](mailto:aramasub@qti.qualcomm.com)  [joels@qti.qualcomm.com](mailto:joels@qti.qualcomm.com)  [garysull@microsoft.com](mailto:garysull@microsoft.com)  [yekuiw@qti.qualcomm.com](mailto:yekuiw@qti.qualcomm.com)  [yan.ye@InterDigital.com](mailto:yan.ye@InterDigital.com) |
| *Source:* | Editors | | |

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

# Executive summary

SHVC, the scalable form of HEVC, codes video data of different qualities into *layers*, and allows subsets of the layers of video data to be extracted and used on their own. Verification tests of SHVC have been conducted to evaluate the amount of savings that can be achieved by SHVC compared to HEVC simulcast, where video data of different qualities are separately coded into independent bitstreams using HEVC. Following an initial set of SHVC verification tests that were the subject of a previous report in February 2016 [1], this document reports the results for supplemental verification testing of SHVC compression performance to test an additional use case.

A subjective evaluation was conducted to compare the SHVC Scalable Main 10 profile to HEVC *simulcast* using the HEVC Main 10 profile, respectively, based on the compression performance of reference software developed as part of the SHVC development process. The supplemental tests described in this report covered the case with a standard dynamic range (SDR) base layer (in a BT.2020 “container” representation) and a high dynamic range (HDR) enhancement layer (using the SMPTE ST 2084 “PQ” transfer characteristics with BT.2020 colour primaries), with both of these using a representation with BT.2020 colour primaries. This contrasts with the previous verification test of SHVC with an HDR enhancement layer, in which the SDR base layer had used a BT.709 “container” representation.

The tests were performed using a test methodology described in this report. The results confirmed that the tested SHM 11.0 reference software encodings using the SHVC Scalable Main 10 profile achieved the same subjective quality as the HEVC simulcast encodings using the Main 10 profile while requiring on average approximately 45% less bitrate for the case when both the SDR base layer and the HDR enhancement layer are coded using the BT.2020 colour primaries (with BT.2020 transfer characteristics in the base layer and SMPTE ST 2084 “PQ” transfer characteristics in the enhancement layer).

# Test sequences and conditions

## Test sequences

The sequences used in the supplemental SHVC verification tests are listed in Table 1. Also provided in the table are the sequence characteristics, including the sequence name, frame rate, and the sequence length used to encode each sequence. An encoding prediction structure configuration enabling random access (RA) was used as further described in Annex B. The sequences in the test were coded using the Scalable Main 10 profile or the Main 10 profile. The test sequences have the following characteristics:

* SDR container format: BT.2020 (with conventional SDR transfer characteristics)
* HDR container format: BT.2020 colour primaries with SMPTE ST 2084 “PQ” transfer characteristics
* Gamut: BT.709 or P3D65, depending on the content
* Original (not coding) colour format: RGB 4:4:4 for HDR and YCbCr 4:2:0 for SDR
* Coding format: 10 bit 4:2:0
* Spatial resolution: 1920x1080

Table 1 Test sequences used for the supplemental SHVC verification tests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sequence | Frame rate (fps) | Gamut | Length (frames) | Length (seconds) |
| Market3 | 50 | BT.709 | 400 | 8 |
| ShowGirl | 25 | P3D65 | 339 | 13.56 |
| EBU\_06\_Starting | 50 | BT.709 | 500 | 10 |
| EBU\_04\_Hurdles | 50 | BT.709 | 500 | 10 |
| WarmNight | 24 | P3D65 | 361 | 15 |
| BalloonFestival | 24 | BT.709 | 240 | 10 |
| GarageExit | 24 | P3D65 | 288 | 12 |



**Market ShowGirl**

****

**EBU\_06\_Starting EBU\_04\_Hurdles**



**WarmNight BalloonFestival**



**GarageExit**

Figure 1: Snapshots illustrating video test sequences used in the test

The EBU\_06\_Starting and EBU\_04\_Hurdles sequences were originally shot in 100 fps and were played out in 50 fps (slow motion), which follows the visual testing practice used in previous HDR AhG of MPEG and JCT-VC. The SDR sequences for Market3, ShowGirl, and WarmNight sequences are manually graded and the SDR sequences for the BalloonFestival, GarageExit, EBU\_06\_Starting, and EBU\_04\_Hurdles sequences are automatically graded. The colour gamut of the Market3, EBU\_06\_Starting, EBU\_04\_Hurdles, BalloonFestival and GarageExit sequences is the same for the base and enhancement layers. For the WarmNight and ShowGirl sequences, the colour gamut of the base layer sequences is BT.709.

## Test conditions

The performance of SHVC encodings was compared to the performance of HEVC simulcast encodings for a wide range of bitrates that cover a wide range of subjective qualities.

In HEVC and its SHVC extensions, as in most international standards for video coding, the encoding method is left outside the scope of the video coding standard. Only the format of the bitstream syntax and the decoding process are standardized. (Encoder pre-processing, decoder post-processing, display adaptation, and recovery from data losses and corruption are also left outside the scope of the standard.) This particularly allows encoder designers the freedom to develop their own encoding algorithms, while ensuring that *interoperability* for decoding is maintained. Nevertheless, when trying to assess the compression capability of a standard, it is necessary for some particular encoding algorithm(s) to be selected to reasonably represent its capability. Moreover, when comparing different coding schemes under well-controlled circumstances, it is generally helpful to use comparable encoding techniques and configurations for the coding schemes that will be compared. In this instance, a recent version of the reference software known as the SHM (Scalable HEVC model) [2], which was developed as part of the work on developing the SHVC standard itself, was used for this verification test. The HM software developed as part of the development of the HEVC standard was used to generate the HDR simulcast stream.

The SHVC encodings were performed using software based on the SHM 11.0 reference software, and the HDR simulcast encodings were performed using software based on HM-16.7 reference software. For SHVC, the encoding of each test sequence was configured to satisfy an additional condition, that is, the enhancement bitrate ratio of the EL relative to the BL was kept roughly around 10%–20%. Each test sequence was encoded at four bitrates. For each video test sequence and each bitrate point, encoding was performed based on the constant-QP configuration as defined in the well-established SHVC common test conditions [3]. For SHVC, the EL and BL QP values were selected such that the enhancement bitrate ratio of the EL relative to the BL was kept roughly around 10%–20%. For HEVC simulcast, the same BL QP values as used in the corresponding SHVC encodings were used, and the EL QP values were chosen such that the quality of the simulcast “EL” and SHVC EL have roughly the same subjective quality. Further information on the detailed coding configurations can be found in Annex B.

# Test environment and methodology

The test procedure for the formal subjective evaluation considered two main requirements:

* to be as reliable and effective as possible in verifying the performance in terms of subjective quality (and therefore adhering to the existing recommendations on assessment methods for the testing of subjective video quality for multimedia applications [4]);
* to take into account the evolution of technology and laboratory set-up requirements oriented to the adoption of appropriate Flat Panel Display (FPD) and video server systems as video recording and playing equipment.

Therefore, one of the test methods described in [4], specifically the Degradation Category Rating (DCR) method, was used. Some modifications were applied, in relation to the kind of displays and the video recording and play-back equipment. Further information on the DCR method can be found in Annex A.

In subjective viewing sessions, two test cases were compared: the HEVC simulcast coded enhancement layer HDR video (labelled as P01) and the SHVC coded enhancement layer HDR video (labelled as P02). The video corresponding to the enhancement layer was displayed as the “coded” video in Figure 4 in Annex A; the video corresponding to the base layer was not viewed and compared, as they were identical between the SHVC and HEVC simulcast test cases. Further information about the test laboratory set-up can be found in Annex A.

The video test sequences were renamed as shown in Table 2. The sequences were renamed as “P0xS0xCXRn” using the following convention:

* P0x identifies the test case, with P01 being the HEVC simulcast test case and P02 being the SHVC test case.
* S0x identifies the sequence number, from S01 to S07.
* Rn identifies the coded bitrate; four rates, R1 to R4, were used to identify bitrates from the highest to the lowest.

Table 2 Numbering of test sequences according to viewing session

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sequence no. | Sequence name | Frame rate | Frame rate displayed | Coding config. |
| P0xS01C4Rn | Market | 50 fps | 50 fps | RA |
| P0xS02C4Rn | ShowGirl | 25 fps | 25 fps | RA |
| P0xS03C4Rn | EBU\_06\_Starting | 50 fps | 50 fps | RA |
| P0xS04C4Rn | EBU\_04\_Hurdles | 50 fps | 50 fps | RA |
| P0xS05C1Rn | GarageExit | 24 fps | 25 fps | RA |
| P0xS06C1Rn | BalloonFestival | 24 fps | 25 fps | RA |
| P0xS07C2Rn | WarmNight | 24 fps | 25 fps | RA |

# Test results

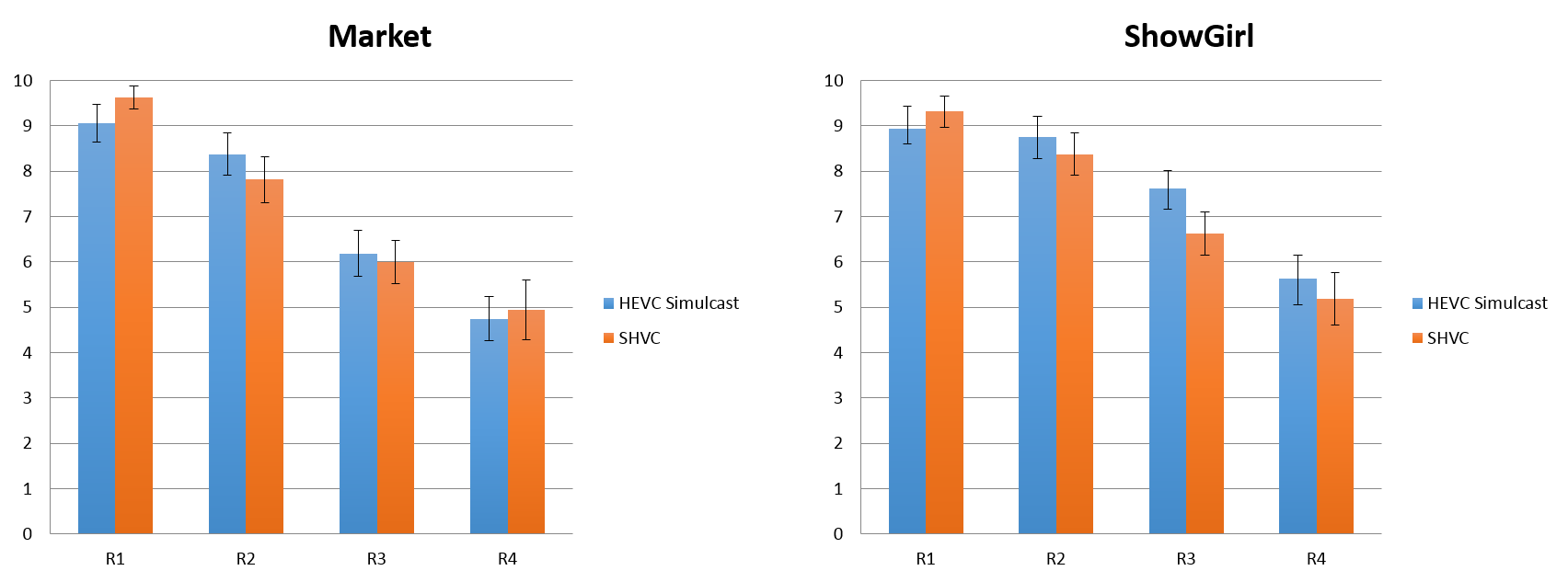
The eleven-grade mean-opinion score (MOS) scale ranging from 0 to 10 was used for the plots in this section. They may be interpreted as shown in Table 3.

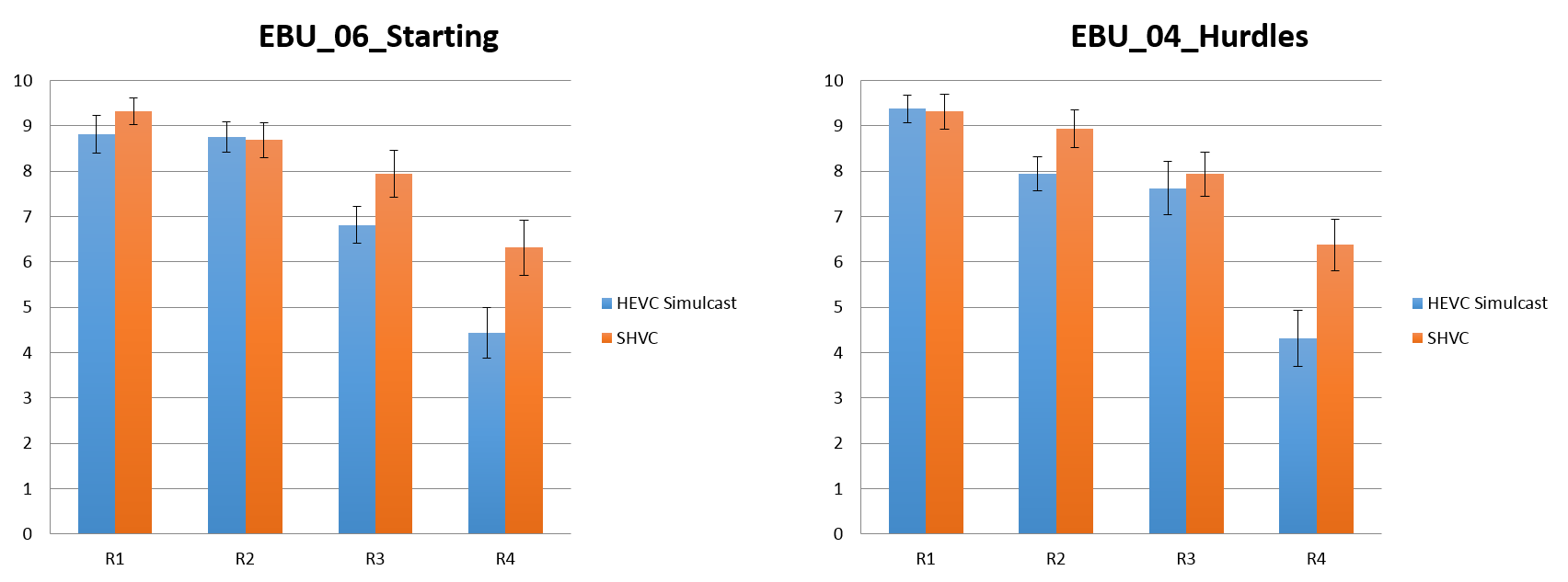
Table 3 Interpretation of MOS score values

|  |  |
| --- | --- |
| 10 - | The number 10 denotes a quality of reproduction that is perfectly faithful to the original. No further improvement is possible. |
| 9 - | Excellent |
| 8 - |  |
| 7 - | Good |
| 6 - |  |
| 5 - | Fair |
| 4 - |  |
| 3 - | Poor |
| 2 - |  |
| 1 - | Bad |
| 0 - | The number 0 denotes a quality of reproduction that has no similarity to the original. A worse quality cannot be imagined. |

## Subjective viewing results

The subjective viewing results were collected and the mean and confidence interval (CI) of the MOS values for each test point are presented below in Figure 2.





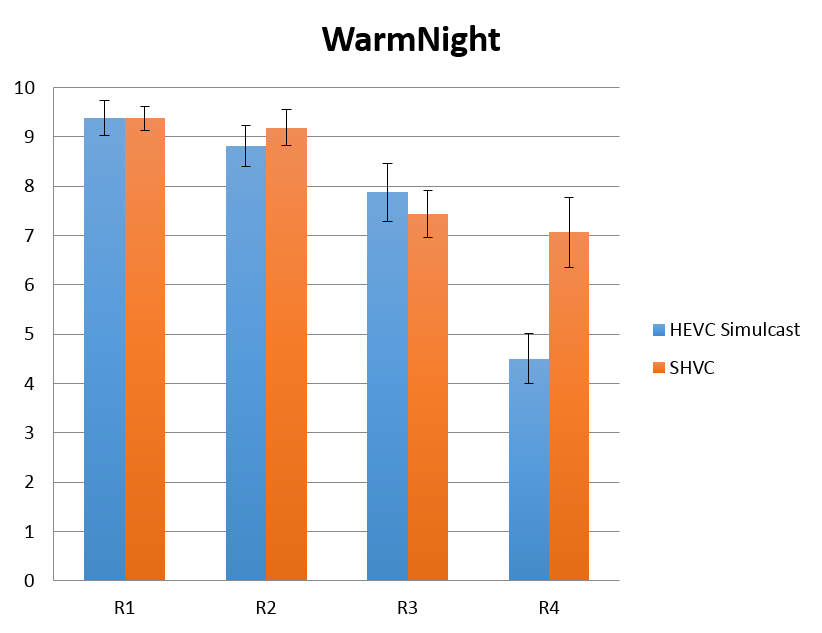
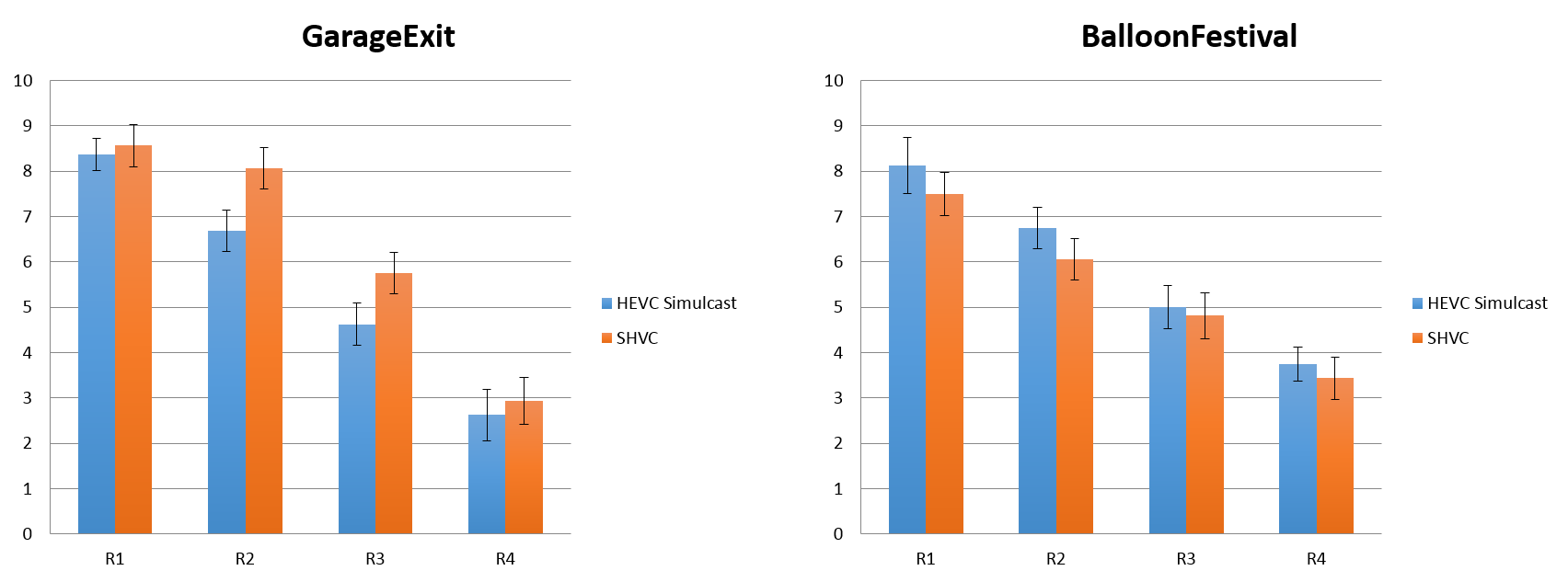
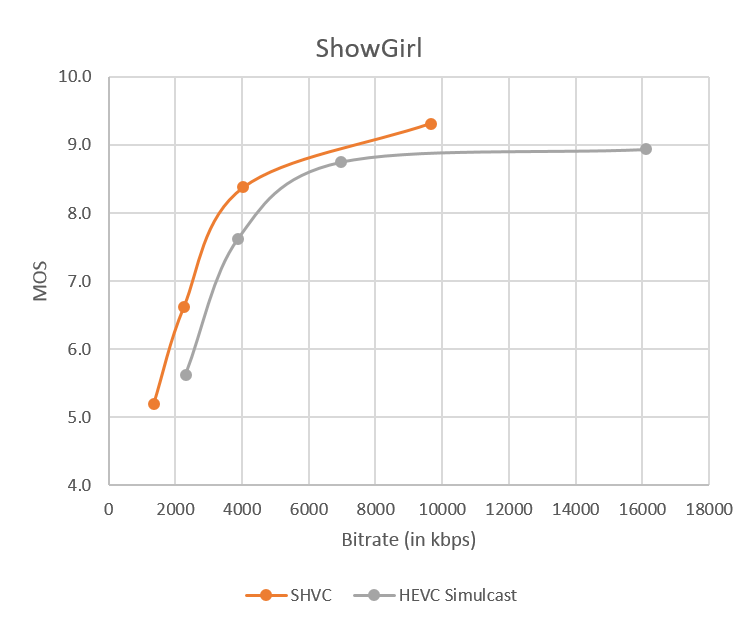
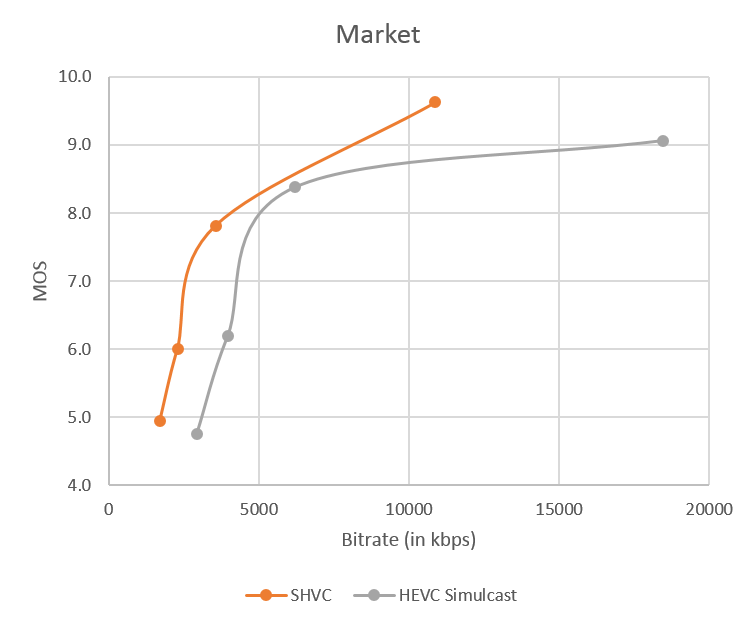
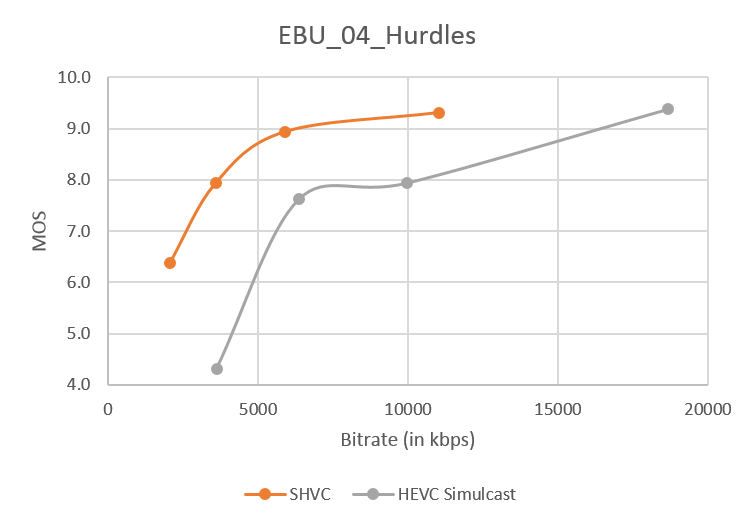
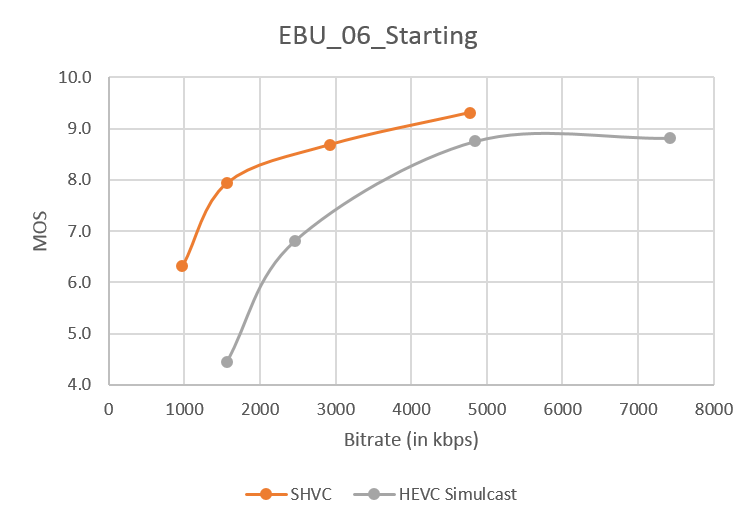
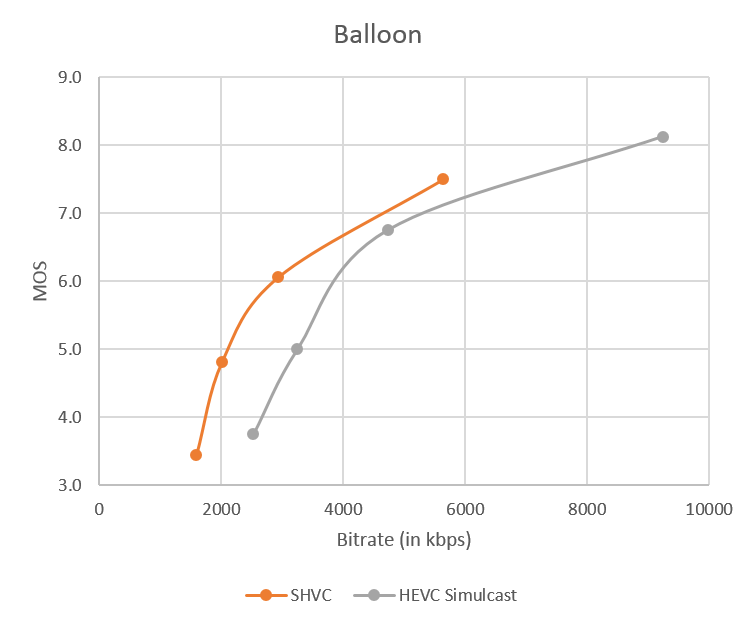
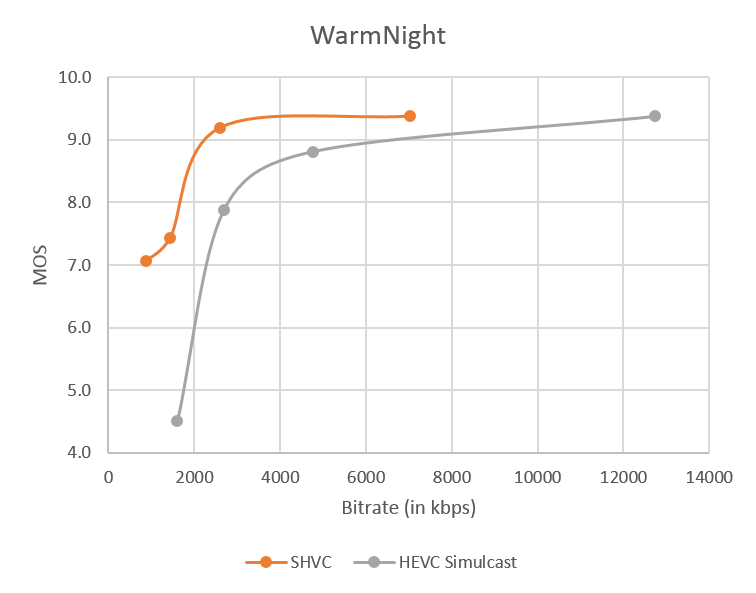


Figure 2: Subjective viewing results for SHVC and HEVC simulcast

## Plots of MOS vs. bitrates

In this section, plots of bitrate vs. mean MOS are provided in Figure 3 for all test sequences used in the supplemental SHVC verification test. All plots show that the bitrate vs. MOS curves of the SHVC test points are located substantially to the left of the corresponding bitrate vs. MOS curves of the HEVC simulcast test points. This confirms that the SHVC encodings achieved a substantial bitrate reduction over HEVC simulcast at the same subjective quality.

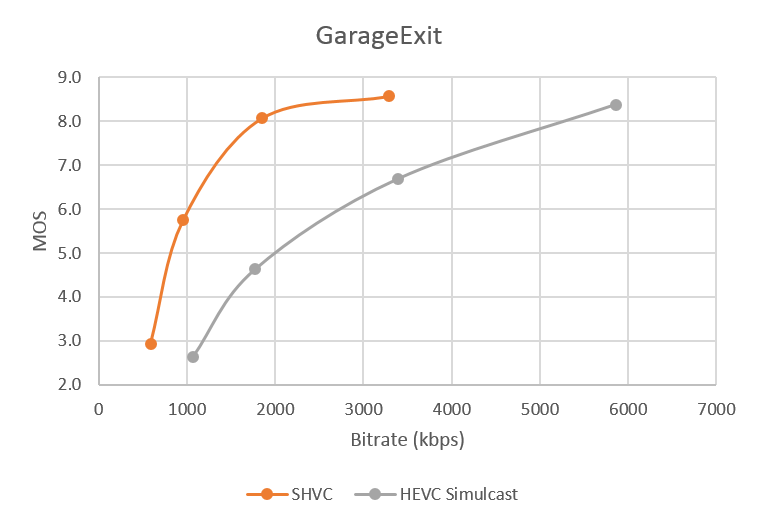


Figure 3: Bitrate vs. MOS curves for all test sequences

The average bitrate savings of SHVC compared to HEVC simulcast for each sequence were computed from the MOS vs. bitrate data to further quantify the savings achieved. Table 4 shows the MOS Bjøntegaard-Delta-rate (BD-rate) for each sequence. BD-rate measures as described in [5] were used with MOS scores taking the place of the peak signal-to-noise ratio (PSNR) values that have been typically used with BD-rate measurements, with negative numbers indicating percentage of rate reduction at the same MOS quality. When considering (BL + EL) bitrates, the measured relative bitrate savings achieved by SHVC compared to HEVC simulcast ranged from 23.5% to 60.5%.

Table 4 MOS BD-rate savings measurements

|  |  |
| --- | --- |
| **Test sequence** | **BD-rate using MOS (BL+EL)** |
| Market | −28.2% |
| BalloonFestival | −37.9% |
| EBU\_06\_Starting | −49.3% |
| EBU\_04\_Hurdles | −23.5% |
| ShowGirl | −59.0% |
| WarmNight | −51.2% |
| GarageExit | −60.5% |
| **Average** | **−44.2%** |

# Conclusion

The subjective test results show that, based on reference software encodings, the SHVC and HEVC simulcast encodings have similar subjective quality (that is, with overlapping confidence intervals) for 23 out of the 28 test points. For the remaining test points, the SHVC encodings provide better subjective quality. For all test points, the SHVC bitrates are substantially lower than the HEVC simulcast bitrates.

By applying the MOS BD-rate measurement on the results of the subjective test, it is estimated that SHVC achieves the same subjective quality as HEVC simulcast while requiring on average 45% less bitrate, when both the base and enhancement layer video are coded in BT.2020 colour primaries, with the base layer being SDR video in a BT.2020 container and the enhancement layer being HDR video using the SMPTE ST 2084 “PQ” transfer characteristics function.

Therefore, it is concluded that the project objective of achieving a substantial bitrate reduction over HEVC simulcast has been met by SHVC for the additional use case tested in this supplemental verification test.

# Acknowledgements

The JCT-VC wishes to thank all the organizations and individuals who contributed to the supplemental SHVC verification test, including:

* Cable Television Laboratories, Technicolor, Dolby Laboratories, Hochschule der Medien Stuttgart, Digital Cinema Initiatives, LLC (DCI), and European Broadcasting Union for providing the test sequences.
* Qualcomm Inc. and InterDigital Communications Inc. for providing the resources to prepare the test material.
* Dr. Vittorio Baroncini (MPEG Test Chairman) for his guidance and coordination of the subjective test.

# References

1. [JCTVC-W1004](http://phenix.it-sudparis.eu/jct/doc_end_user/documents/23_San%20Diego/wg11/JCTVC-W1004-v1.zip), “Verification Test Report for Scalable HEVC (SHVC),” Feb. 2016.
2. JCTVC SHM11.0 reference software, available at <https://hevc.hhi.fraunhofer.de/svn/svn_SHVCSoftware/tags/SHM-11.0>.

1. [JCTVC-Q1009](http://phenix.it-sudparis.eu/jct/doc_end_user/current_document.php?id=9106), “Common SHM test conditions and software reference configurations,” April 2014.
2. International Telecommunication Union – Telecommunication Standardization Sector; [Recommendation ITU-T P.910](http://www.itu.int/rec/T-REC-P.910) *Subjective video quality assessment methods for multimedia applications*.
3. Gisle Bjøntegaard, "Improvements of the BD-PSNR model", ITU-T SG16/Q6, 35th VCEG meeting document [VCEG-AI11](http://wftp3.itu.int/av-arch/video-site/0807_Ber/VCEG-AI11.zip), Berlin, Germany, July, 2008.

1. [JCTVC-X0035](http://phenix.it-sudparis.eu/jct/doc_end_user/current_document.php?id=10493), “SHM encoder improvements for HDR verification testing”, May-June, 2016.

1. [JCTVC-W1017](http://phenix.it-sudparis.eu/jct/doc_end_user/current_document.php?id=10471), “Conversion and Coding Practices for HDR/WCG Video, Draft 1”, Feb. 2016
2. [JCTVC-W1020](http://phenix.it-sudparis.eu/jct/doc_end_user/current_document.php?id=10487), “Updated Common Test Conditions for HDR/WCG Video Coding Experiments,” Feb. 2016
3. Pseudo Isochromatic Plates, engraved and printed by The Beck Engraving Co., Inc., Philadelphia and New York, United States.
4. Kirk, R.E., Experimental Design – Procedures for the Behavioural Sciences, 2nd Editions, Brooks/Cole Publishing Co., California, 1982.

# Annex A: Testing procedure

*Test method*

The test method adopted for this evaluation was DCR (Degradation Category Rating) [4].

*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.

This method is used under the schema of evaluation of the quality (and not of the impairment); for this reason, a quality rating scale made of 11 levels was adopted, ranging from "0" (lowest quality) to "10" (highest quality). The test will be held in different laboratories located in countries speaking different languages. This implies that it is better not to use categorical adjectives (e.g. excellent good fair etc.) to avoid any bias due to a possible different interpretation by naive subjects speaking different languages.

The structure of the Basic Test Cell (BTC) of DCR method was made by using two consecutive presentations of the video clip under test; at first the original version of the video clip was displayed, immediately afterwards the coded version of the video clip was presented, and this is repeated once; then a message was displayed for 5 seconds asking the viewers to vote (see Figure 4).

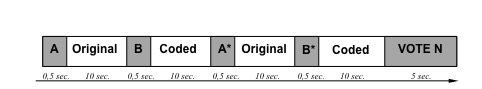


Figure 4: DCR BTC

The presentation of the video clips was preceded by a mid-grey screen displaying for one second.

*How to express the visual quality opinion with DCR*

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

The scoring sheet for a DCR test was made of a section for each BTC; each section was made of a column of 11 vertically arranged boxes, each associated with a number from 0 to 10 (see Figure 5).

The viewers were required to put a check mark on one of the 11 boxes; checking the box "10" if the subject expresses an opinion of "best" quality, while checking the box "0" if the subject expresses an opinion of the "worst” quality.

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

Scoring sheet Class C IVC english

Figure 5: Example of DCR test method scoring sheet

*Training and stabilization phase*

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

For this purpose, each subject was trained by means of a short practice (training) session.

As a testing principle, the video material used for the training session was required to be different from those of the test, but the impairments introduced by the coding were kept, as much as possible, similar to those in the test.

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

The scores of the stabilization phase were discarded. Consistency of the behaviour of the subjects was checked by inserting in the session a BTC in which the original is compared to the original.

*The laboratory set-up*

The test was done using a prototype SIM2 HDR display for which the manufacturer declared a peak luminance close to 9,000 nits in the “HDR-mode”, over the 40% of the whole screen.

A screen luminance uniformity measurement was done by means of an X-Rite luminance meter (model i1Display Pro), getting a mean luminance value around 6,400 nits, with a variation of around 2% across the 9 quadrants of the screen (up-left, up-center, up-right, center-left, center-center, center-right, down-left, down-center, down-right). The measurement was done with the display in the “non HDR” mode.

All the test signals (included the caption messages: A B A\* B\* and VOTE N) were sent to the monitor through the LogLuv DVI input. The video signals were all in AVI LogLuv file format, while the caption messages were bitmap fixed images, whose presence on the screen was set by means of a script of the MUP (Multimedia Universal Player) HDR player.

The laboratory for a subjective assessment was set up according to [4], except for the selection of the display and the video play-out server.

For video sequences with 24 and 30 frames per second, they were displayed at 25 frames per second. For video sequences at 50 frames per second, they were displayed at 50 f frames per second. The frame rate at which each of the S01 to S07 test sequences was displayed is provided in Table 2.

*Viewing distance*

The viewing distance was 2.5 H, where H is equal to the height of the screen. Two subjects were seated in front of the SIM2 monitor at the same time.

*Viewing environment*

The test laboratory has to be carefully protected from any external visual or audio pollution.

Internal general light has to be low (just enough to allow the viewing subjects to fill out the scoring sheets) and a uniform light has to be placed behind the monitor, in a way no direct light hits the viewing subjects seated in front of the screen; a LED lamp (AwoX SmartLight White; <http://store.awox.com/fr/p_awox_product.aspx?i=270244>) , dimmed to around 25 lumen and set for a colour temperature of 6500°, was placed behind the monitor to illuminate the wall. No other light source was on, ceiling, floor and walls of the laboratory were made of non-reflecting material.

*Overall test effort and subjects’ involvement*

The duration of the test depended on the number sequences tested in each category/resolution assigned to the test laboratory; in any case each viewing session did run for more than 20 minutes and the same viewing subject did not participate in the test run for more than six hours in total. The same subject was also required not to be enrolled for two consecutive days. Young human subjects, equally distributed in gender, between the age of 18 to 30, and highly preferably among university students of scientific faculties, were hired. Viewing subjects were compensated for their participation in the testing activities (with compensation in money or services).

*Statistical analysis and presentation of the results*

The data collected from the score sheets, filled out by the viewing subjects, was stored in Excel spreadsheets.

For each coding condition the Mean Opinion Score (MOS) and associated Confidence Interval (CI) values were computed in the spreadsheets.

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

From the “raw” data, subject reliability was calculated, and the method used to assess subject used criteria for subjective reliability as given in [9] and [10].

As an example, the reliability of a subject could be determined by computing the correlation index between each score provided by a subject to the general MOS value assigned for that test point; in this regard a correlation index equal or superior to 0.75 (computed making the mean of all the correlation values) could be considered as valid for the acceptance of the subject.

# Annex B: Additional information on encoding configurations

The following encoding configuration parameters are used:

* Fixed QP
  1. 4 bitrate points per sequence covering the whole MOS range as much as possible
* Bit depth: 10 bits
* Coding structure
  1. Random access (RA)
     1. IRAP picture period at approximately 1 second
     2. Hierarchical B coding structure
* Bitstreams and other settings:
  1. Anchor 3.2 configuration bitstreams generated as described in Annex C.
  2. SHM configuration bitstreams generated as described in Annex C.

# Annex C: Procedure applied to produce test bitstreams

The following steps describe how to produce the anchor 3.2 configuration bitstreams:

1. Download HDRTools using SVN from <http://wg11.sc29.org/svn/repos/Explorations/XYZ/HDRTools/tags/0.10/>  
   (Note: HDRTools is also now available at https://gitlab.com/standards/HDRTools.git.)
2. Download the updated CTC package. Go to http://wg11.sc29.org/ using a browser. Click on “MPEG-content” and navigate to [http://wg11.sc29.org/content/Explorations/HDR/CEs\_Software\_113/CE1/](http://wg11.sc29.org/content/Explorations/HDR/CEs_Software_113/CE1/updatedCTCsPackage_20160127.zip). Download updatedCTCsPackage\_20160127.zip.
3. Update HDRTools and its config files with HDRTools-0.10-src.patch and HDRTools-0.10-cfg.patch from the updated CTCs package.
4. Convert source sequences to Y’CbCr using the patched HDRTools.
5. Compile HM\_anchor3.2 from the updated CTC package and encode the sequences using config files encoder\_randomaccess\_main10\_classAGH\_vui\_lumaQP\_r1.cfg for the BT.709 gamut sequences and encoder\_randomaccess\_main10\_classBCD\_vui\_lumaQP\_r1.cfg for the P3 gamut sequences.

The following steps describe how to produce the SHM configuration bitstreams:

1. Download HDRTools using SVN from <http://wg11.sc29.org/svn/repos/Explorations/XYZ/HDRTools/tags/0.10/>
2. Download the SHM software patch of JCTVC-X0035 [6] and apply it to SHM-11.0.
3. Update HDRTools and its config files with HDRTools-0.10-src.patch and HDRTools-0.10-cfg.patch from the updated CTCs package
4. Convert HDR source sequences (enhancement layer) to Y’CbCr using the patched HDRTools.
5. Convert source SDR BT.709 gamut sequences (base layer) using the patched HDRTools and first the config file HDRConvert\_Step1\_2\_YCbCr420ToRGBprime444.cfg and then the config file HDRConvert\_Step3\_RGB444toYUV420.cfg (as per JCTVC-W0046).
6. Compile the patched SHM-11.0 encode the sequences using config files encoder\_SRA\_m10\_vui.cfg and layers\_vui\_sei.cfg included in JCTVC-X0035 [6].