**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2015/N15350**

**June 2015, Warsaw, Poland**

|  |  |
| --- | --- |
| **Source:** | Requirements |
| **Status:** | Approved |
| **Title:** | Test Results of Call for Evidence (CfE) for HDR and WCG Video Coding |
|  |  |

**Abstract**

This document provides the results of test conducted by MPEG Test group to evaluate the visual quality of the responses received corresponding to the Call for Evidence (CfE) associated with technology for possible extensions of High Efficiency Video Coding (HEVC) standard for High Dynamic Range (HDR) and Wide Colour Gamut (WCG) video coding applications. These tests were conducted by two laboratories – one in Lausanne and other in Rome.

# Introduction

Since the completion of the first edition of the High Efficiency Video Coding (HEVC) standard, several key extensions of its capabilities have been developed to address the needs of broader range of applications. Recognizing the increasing interest in High Dynamic Range (HDR) and Wide Color Gamut (WCG) applications, the Moving Picture Experts Group (MPEG) released in February 2015 a Call for Evidence (CfE) for HDR and Wide Color Gamut (WCG) video coding [A. Luthra, E. Francois and W. Husak, “Call for Evidence (CfE) for HDR and WCG Video Coding”, ISO/IEC JTC1/SC29/WG11/N15083, February 2015, Geneva, Switzerland]. The purpose of this CfE was to explore whether the coding efficiency and/or the functionality of current HEVC standard can be significantly improved for HDR and WCG content.

The CfE covered different HDR and WCG applications, including backward compatibility with existing Standard Dynamic Range (SDR) decoders and displays. Several anchor sequences were produced using HEVC Main 10 and Scalable Main 10 Profiles as reference sequences as described in the CfE. In total, nine responses were received. The results of the subjective evaluation tests conducted on those responses by EPFL in Lausanne are reported in m36728 and attached here in Annex A. The results of the subjective evaluation test conducted on those responses by Mr. Vittorio Baroncini (MPEG Test Chair) in Rome are reported in m36836 and attached here in Annex B.

# Conclusion

EPFL and Rome tests show that there are submitted technologies demonstrating statistically significant improvements over the anchors.

# Acknowledgement

We are thankful to Apple, Arris, BBC, Dolby, Ericsson, FastVDO, InterDigital, MovieLabs, NGCodec, Philips, Qualcomm, Technicolor and University of Warwick/goHDR for responding to HDR and WCG CfE.

We are thankful to EPFL for hosting AHG meeting as well as conducting the subjective tests in Lausanne. We are thankful to MPEG Test Chair, Vittorio Baroncini, for his guidance and conducting the subjective tests in Rome. We are thankful to Dolby, Technicolor and SIM2 for providing displays to help in conducting our work.

Annex A

(This page left blank)

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2015/m36728**

**July 2015, Warsaw, Poland**

|  |  |
| --- | --- |
| **Source** | **Ecole Polytechnique Fédérale de Lausanne (EPFL)** |
| **Status** | **Input document** |
| **Title** | **Results of HDR CfE subjective evaluations conducted at EPFL** |
| **Author** | **Philippe Hanhart, Martin Rerabek, and Touradj Ebrahimi** |

# Abstract

This contribution reports the details and results of the subjective evaluations conducted at EPFL to evaluate the responses to the Call for Evidence (CfE) for HDR and WCG Video Coding. In total, nine submissions, five for Category 1 and four for Category 3a, were compared to the HEVC Main 10 Profile based Anchor. The subjective evaluations were conducted following the methodology presented during the AhG meeting in Lausanne in May. Five HDR video contents, compressed at four bit rates by each proponent responding to the CfE, were used in the subjective evaluations. Subjective results show that, on some contents, there is evidence that some proponents achieve statistically significantly better visual quality than the Anchor.

# Introduction

Since the completion of the first edition of the High Efficiency Video Coding (HEVC) standard, several key extensions of its capabilities have been developed to address the needs of an even broader range of applications. Recognizing the rise of High Dynamic Range (HDR) applications and the lack of a corresponding video coding standard, the Moving Picture Experts Group (MPEG) has released in February 2015 a Call for Evidence (CfE) for HDR and Wide Colour Gamut (WCG) video coding [1]. The purpose of this CfE is to explore whether the coding efficiency and/or the functionality of HEVC can be significantly improved for HDR and WCG content.

The CfE covers different applications, e.g., backward compatibility with existing Standard Dynamic Range (SDR) content, as well as different categories, e.g., normative and non-normative changes to HEVC. In total, eight companies or consortia of different companies and one university responded to the CfE and submitted responses to one or more of the different categories. Initially, only responses to categories 1, 2b, 3a, and 3b were planned to be tested through formal subjective evaluations. Based on the number of responses, it was further agreed during the AhG meeting in Lausanne in May that only responses to categories 1 and 3a would be tested in the formal subjective evaluations, while the remaining categories would be evaluated otherwise. Therefore, the results reported in this contribution cover only the five Category 1 submissions, i.e., P11, P12, P13, P14, and P22, and the four Category 3a submissions, i.e., P31, P32, P33, and P34.

# Subjective evaluation

## Dataset

The dataset used for the subjective evaluation tests consists of five HD resolution HDR video sequences, namely, Market3, AutoWelding, ShowGirl2, WarmNight, and BalloonFestival. Each video sequence was cropped to 950 x 1080 pixels, so that the video sequences were presented side-by-side with a 20-pixels separating black border. Each video sequence was displayed at 24 fps, which is the native frame rate of the display used in the experiments (see Section 2.2), and cut to 240 frames, which corresponds to 10 seconds. Note that the Market3 sequence was played at a slower frame rate than the original content (50 fps). This solution was evaluated as visually more pleasant than playing every other frame, which created temporal distortion. The coordinates of the cropping window and selected frames are given in Table 1.

Table 1: HDR test sequences used in the subjective evaluations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Seq** | **Sequence name** | **Cropping window** | **Selected frames** |
| S02 | Market3 | 970-1919 | 0-239 |
| S03 | AutoWelding | 600-1549 | 162-401 |
| S05 | ShowGirl2 | 350-1299 | 94-333 |
| S07 | WarmNight | 100-1049 | 36-275 |
| S08 | BalloonFestival | 0-949 | 0-239 |

The data was stored in uncompressed 16 bit TIFF files, in 12 bit non-linearly quantized (using Dolby PQ EOTF) RGB signal representation, using the SDI data range (code values from 16 up to 4076) and BT.2020 RGB colour space.

The side-by-side video sequences were generated using the HDRMontage tool from the HDRTools package [3].

## Test environment

The experiments were conducted at the Multimedia Signal Processing Group (MMSPG) test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU-R [4]. The test room is equipped with a controlled lighting system of a 6500 K color temperature. The color of all background walls and curtains in the room is mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors. In the experiments, the luminance of the background behind the monitor was about 20 cd/m2. The ambient illumination did not directly reflect off of the display.

To display the test stimuli, a full HD (1920 × 1080 pixels) 42'' Dolby Research HDR RGB backlight dual modulation display (aka Pulsar) was used. The monitor has the following specifications: full DCI P3 color gamut, 4000 cd/m2 peak luminance, low black level (0.005 cd/m2), 12 bits/color input with accurate and reliable reproduction of color and luminance.

In every session, three subjects assessed the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of about 3.2 times the picture height, as suggested in recommendation ITU-R BT.2022 [5].

## Methodology

Two video sequences were presented simultaneously in side-by-side fashion. Since only one full HD 1920 × 1080 HDR monitor was available, each video was cropped to 950 × 1080 pixels with 20 pixels of black border separating the two sequences. One of the two video sequences was always the Anchor, with a randomized position on the screen (either on the left or on the right). The other video sequence was the Proponent to be evaluated, at the same (targeted) bit rate as the Anchor.

Subjects were asked to judge which video sequence in a pair (‘left’ or ‘right’) has the best overall quality, considering fidelity of details in textured areas and color rendition. The option ‘same’ was also included to avoid random preference selections.

## Test planning

Before the experiments, a consent form was handed to subjects for signature and oral instructions were provided to explain the evaluation task. A training session was organized to allow subjects to familiarize with the assessment procedure. The same contents were used in the training session as in the test session to highlight the areas where distortions can be visible. Eleven training samples were manually selected by expert viewers. First, two samples, one of high quality and one of low quality, without any difference between left and right, were selected from the AutoWelding sequence. The purpose of these two examples was that subjects could get familiar with HDR content, as this content has both dark and bright luminance levels and fast luminance temporal changes, and see the extreme levels of quality observed in the test material. Then, one sample from AutoWelding with large visible difference was presented to illustrate the main differences that can be observed between the left and right video sequences, i.e., loss of texture/details and color artifacts. Finally, for each of the remaining contents, two samples were presented (one example with large difference and one example with small differences) in the following order: Market3, BalloonFestival, ShowGirl2, and WarmNight. The training materials were presented to subjects exactly as for the test materials, thus in side-by-side fashion.

The overall experiment was split into 6 test sessions. Each test session was composed of 30-31 basic test cells (BTC), corresponding to approximately 14 minutes each. To reduce contextual effects, the stimuli orders of display were randomized, whereas the same content was never shown consecutively. The test material was randomly distributed over the six test sessions.

Each subject took part to exactly three sessions. Three dummy pairs, whose scores were not included in the results, were included at the beginning of the first session to stabilize the subjects' ratings. Between the sessions, the subjects took a 14-minute break.

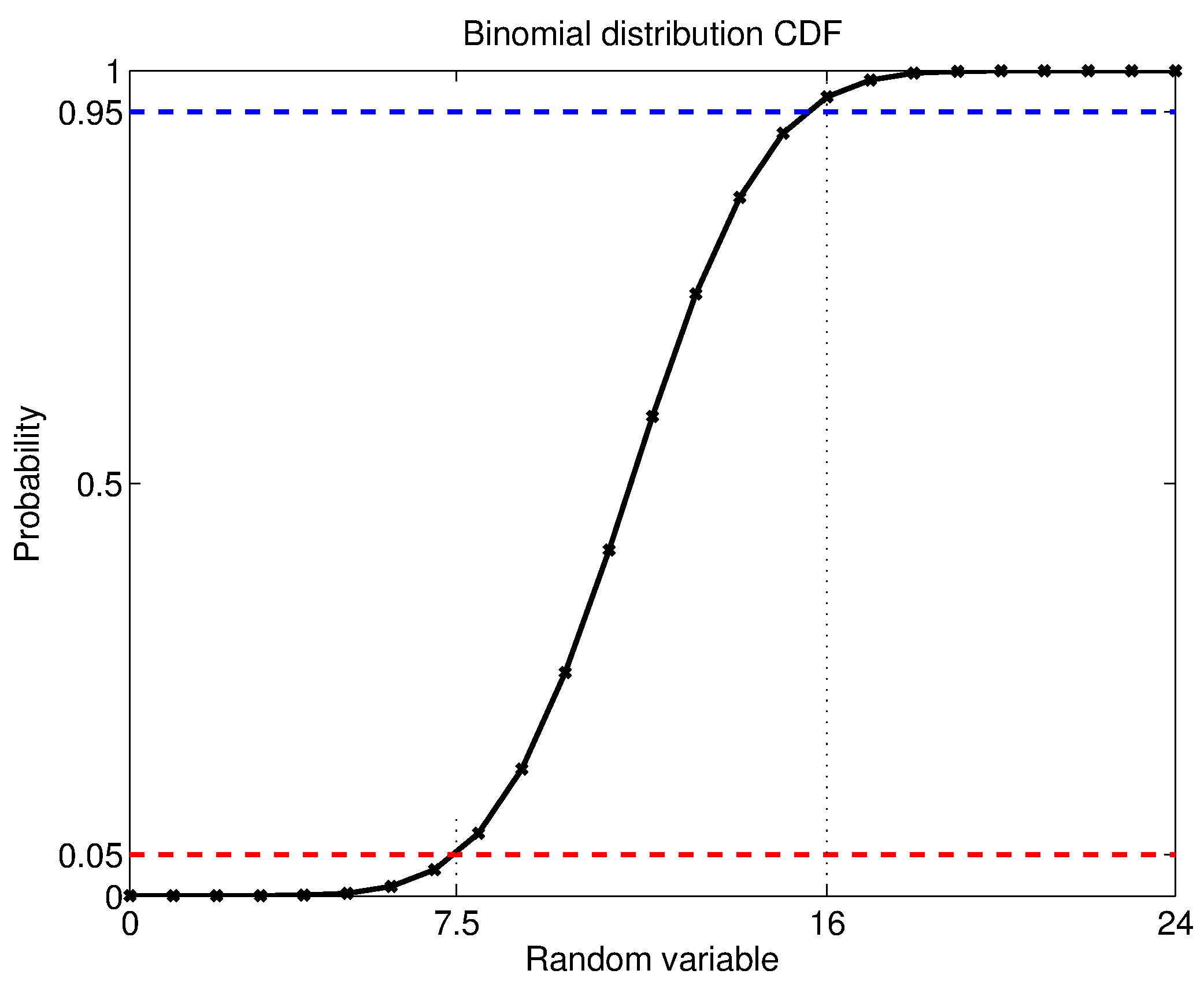
A total of 48 naïve subjects (16 females and 32 males) took part in the experiments, leading to a total of 24 ratings per test sample. Subjects were between 18 and 49 years old with an average and median of 25.3 and 24 years of age, respectively. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

# Statistical analysis

No outlier detection was performed on the raw scores, since there is no international recommendation or a commonly used outlier detection technique for paired comparison results.

For each test condition, i.e., combination of content, algorithm, and bit rate, the winning frequency of the Anchor, , winning frequency of the Proponent, , and tie frequency, , are computed from the obtained subjective ratings. Note that , where is the number of subjects. To compute the preference probability of selecting the proponent version over the Anchor, , ties are considered as being half way between the two preference options: .

To determine whether the visual quality difference between the proponent and the Anchor is statistically significant, a statistical hypothesis test was performed. As ties are split equally between the two preference options, the data roughly follows a Bernoulli process , where is the number of subjects and is the probability of success in a Bernoulli trial and was set to 0.5, considering that, *a priori*, the Anchor and proponent have the same chance of success. Figure 1 shows the cumulative distribution function (CDF) for Binomial distribution with and . The CDF is used to determine the critical region for the statistical test.



*Figure 1 – Cumulative distribution function for Binomial distribution with and .*

To determine whether the proponent provides statistically significant results, a one-tailed binomial test was performed at 5% significance level with the following hypotheses:

H0: Proponent is equal or worse than Anchor

H1: Proponent is better than Anchor

In this case, the critical region for the preference probability over Anchor, , is , as the CDF for 16 or more successful trials is above 95% (see Figure 1, ). Therefore, if there are 16 or more votes in favor of the proponent, the null hypothesis can be rejected.

Similarly, to determine whether the proponent provides statistically significantly lower visual quality than the Anchor, a one-tailed binomial test was performed at 5% significance level:

H0: Proponent is equal or better than Anchor

H1: Proponent is worse than Anchor

In this case, the critical region for the preference probability over Anchor, , is , as the CDF for 7.5 or less successful trials is below 5% (see Figure 1, ). Note that the Binomial distribution is not defined for non-integer values, and that extension is usually obtained using the floor function. Therefore, if there are 7.5 or less votes in favor of the proponent, the null hypothesis can be rejected.

# Results and discussions

Figure 2 reports the preference probability of selecting the proponent version over the Anchor for each content separately. Category 1 submissions (P11, P12, P13, P14, and P22) are plotted with plain lines, while Category 3a submissions (P31, P32, P33, and P34) are plotted with dashed lines. Values on or above the horizontal blue dashed line provide statistically significant visual quality superior to the Anchor, while values on or below the horizontal red dashed line provide statistically significant inferior visual quality when compared to the Anchor.

As it can be observed, there is evidence that potential coding technologies can do significantly better than the Anchor, especially for contents Market3 and BalloonFestival. For instance, on content ShowGirl2, proponent P22 provides statistically significant superior visual quality when compared to the Anchor at rates R1 to R3. Improvements can also be observed for proponents P11 and P12. Regarding content WarmNight, proponents P32 and P22 outperform the Anchor for rates R2 to R4. Proponents P31 and P11 also show gains for specific rate points. Finally, for content AutoWelding, proponent P32 provides gain for rate R2 to R4, while proponent P12 is at the limit for rate R1.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

*Figure 2 – Preference probability of selecting the proponent version over the Anchor. For content ShowGirl2, dummy values are used for one proponent (BBC), as the decoded material was not provided for this sequence.*

In general, proponent P32 seems to perform better on dark contents than on bright contents. Regarding P14, wrong colors were observed through the whole test material, probably due to a wrong color transformation, as well as occasional green noise in the table scene on content WarmNight.

Regarding the selection of contents, bright scenes are better to perceive color artifacts, especially in whitish parts, and loss of details and high frequencies, especially in textured areas. Sequences such as ShowGirl2 and Market3 are good for testing HDR compression. On the other hand, sequences with a wide dynamic range and strong luminance temporal changes, such as AutoWelding although good for demonstrating HDR may not be necessarily best to assess HDR compression performance. Dark scenes are important too, as HDR is not only about high brightness, but it might be hard to see the improvements in these sequences, especially if the previous test sequence was bright, due to the adaptation time of the human eye.

Regarding the test methodology, the side-by-side presentation was beneficial to discriminate small differences between the Anchor and proponents. Repetition also helps improving discrimination power. However, if an absolute ranking of the proponents is required, the Anchor should be replaced by the source uncompressed reference and an absolute impairment scale should be used, as in a regular DSIS test.

# Conclusion

In this contribution, the details and results of the subjective evaluations conducted at EPFL to evaluate the responses to the Call for Evidence (CfE) for HDR and WCG Video Coding have been reported. Subjective results show that, on some contents, for some bit rates, there is evidence that a number of proponents achieve statistically significant superior visual quality when compared to the Anchor.

# Acknowledgement

This work has been conducted in the framework of the Swiss SERI project Compression and Evaluation of High Dynamic Range Image and Video, COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET, and FP7 EC EUROSTAR funded Project - Transcoders Of the Future TeleVision (TOFuTV).

# References

1. ISO/IEC JTC1/SC29/WG11, “Call for Evidence (CfE) for HDR and WCG Video Coding,” Doc. N15083, Geneva, Switzerland, February 2015.
2. ISO/IEC JTC1/SC29/WG11, “HDR CfE Subjective Evaluations as EPFL,” Doc. M36168, Lausanne, Switzerland, May 2015.
3. ISO/IEC JTC1/SC29/WG11, “HDRTools: Software updates,” Doc. M35471, Geneva, Switzerland, February 2015.
4. ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures.” International Telecommunication Union (January 2012).
5. ITU-R BT.2022, “General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays.” International Telecommunication Union, August 2012.

Annex B

(This page left blank)

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2015/M36836**

**July 2015, Warsaw, Poland**

|  |  |
| --- | --- |
| **Source** |  |
| **Status** | **Input document** |
| **Title** | **Results of the HDR CfE Subjective Evaluations at Rome** |
| **Author** | **Vittorio Baroncini** |

**Results of the HDR CfE test run in Rome**

# Abstract

This document describes the procedures and the results of the subjective evaluations conducted at Rome to subjectively assess the responses to the Call for Evidence (CfE) for HDR and WCG Video Coding [1]. Nine submissions, five for Category 1 and four for Category 3a, were received and evaluated together with an Anchor obtained encoding the selected video content by means of HEVC Main 10 Profile. The formal subjective evaluation test were done using a side by side version of the DSIS method. The test in Rome assessed four HDR video content (SRC), encoded at four bit rates. The result of the test showed evidence of superiority some submission that showed better visual quality than the Anchor at the same bit rate.

# Introduction

The CfE issued by MPEG was trying to cover different applications, e.g., backward compatibility with existing Standard Dynamic Range (SDR) content, as well as different categories, e.g., normative and non-normative changes to HEVC.

Only two Categories were covered by the 9 submission received by MPEG, and in details 5 in Category 1 and 4 in Category 3a.

The proponents were labeled with a unique code to allow a performance identification.

The five Category 1 submissions, were labeled as P11, P12, P13, P14, and P22, and the four Category 3a submissions were labeled as P31, P32, P33, and P34.

# Subjective evaluation

## Dataset

The Test in Rome was done on a SIM2 display that to properly show the video HDR video files requested them to be converted in the Luv AVI file format.

This led to receive from the Proponents a set of AVI file that reflected the length of the original Source files; as we can see from the table here below the file length was consistently variable in the range from 8 seconds to 15 seconds.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **seq** | **Sequence name** | **fps** | **Original Frames** | **length** | **New cut** |
| S00 | FireEater2 | 25 | 0-199 | 8 sec. | 0-199 |
| S01 | Tibul2 | 30 | 0-239 | 8 sec. | 0-239 |
| S02 | Market3 | 50 | 0-399 | 8 sec. | 0-399 |
| S03 | AutoWelding | 24 | 0-426 | 17,8 sec. | 231-421 |
| S04 | BikeSparklers | 24 | 0-479 | 20 sec. | Remove this |
| S05 | ShowGirl | 25 | 0-338 | 13,5 sec. | 149-338 |
| S06 | StEM\_MagicHour | 24 | 3527 to 3887 | 15 sec. | 84-273 |
| S07 | StEM\_WarmNight | 24 | 6280 to 6640 | 15 sec. | 10-199 |
| S08 | BalloonFestival | 24 | 0-239 | 10 sec | 0-191 |

Furthermore the limited resources in time and laboratory availability in Rome suggested to select for the test to be done there only a subject of the above test sequences, i.e.: S02, S05, S07 and S08. Furthermore, as evident from the above table, the length of the selected sequences was in some case far beyond the usual 10 seconds.

And also it was noted to be too time consuming to edit to 10 seconds the AVI Luv files received by Proponents.

For this reason the test was done keeping the selected sequences at their original length.

Another important aspect was the selection of the cropping area that was done to allow a side by side presentation of the video clips for the side by side test.

The test in Rome were done using a new video player (MUP) able to play out (among the others) AVI Luv video files also in split screen configuration and selecting the offset of the cropping window.

Due to this MUP capability the cropping area was carefully selected, on pixel basis, as shown in the table here below, were the offset value indicates the horizontal position of start of the selected window.

|  |  |  |
| --- | --- | --- |
| **Seq** | **Sequence name** | **Off-set value** |
| S02 | Market3 | 50 |
| S05 | ShowGirl2 | 400 |
| S07 | WarmNight | 0 |
| S08 | BalloonFestival | 125 |

## Laboratory set-up

Rome testing area was done of a room having both floor ceiling made of non reflecting black surfaces; The wall behind monitor was made of black curtains.

A 20 lux light was placed behind the monitor. No other light source was present in the testing area and all the room was protected by external audible pollution as well.

The display was a 6000 cd/m2 peak luminance 47” SIM2 monitorwhose features can be found at the SIM2 web site.

Three subjects were seated in font of the monitor and 2H.



## Methodology

The uncompressed video sequence (SRC) was presented simultaneously in side-by-side fashion together with a Processed Video Sequence (PVS). The two clips (SRC and PVS) were presented twice for each test point; during the first presentation the SRC was always on the left side, while in the second presentation the SRC was presented on the right side of the screen.

This choice was done to minimize any possible variation in color and/or luminance that the SIM2 display might have time by time. The position fo the SRC and PVS on the screen was always known to the viewing subjects.

Subjects were asked express their opinion according to the DSIS protocol, i.e. to answer the question: “if and, in the case, how much the PVS sample is different from the SRC sample”.

This simple question allowed to the naïve human subjects to understand in an easy way what was asked to them and to avoid to bias their opinion to any particular impairment they may note. They were only advised that impairment (i.e. loss of quality) could be in the area of image texture or color representation.

The viewing subjects were accurately trained before running the actual test, by means of a training session in which some sample of the actual test material was presented in away they can familiarize with the kind of quality they are going to watch at and to the scoring procedure.

Scoring was done using an 11 numerical impairment scale, where 10 represented the “no difference” case and 0 was the “almost completely degraded” case

A total of 24 students (6 females and 18 males) between 18 and 24 years, participated to the tests, and were all pre-screened for visual acuity and color vision using Snellen and Ishihara charts, respectively; the pre-screening lead to exclude one subject (due to color blindness), while the processing of the results led to exclude other two subjects showing very low response to the observed stimula, leading to a total of 21 valid viewers.

**DSIS SIDE-BY-SIDE with two repetitions**

The following rules are applied:

* The screen is split in two equal areas: left (L), right (R).
* The screen shows on one half the “reference” (SRC) and on the other half the “processed” (PVS) version of the same video clip.
* The Basic Test Cell of the DSIS SbS method is done of two presentation of the same SCR and PVS pair.
* During the first presentation, announced by the letter “A” on the screen, the SRC is on the left side and the PVS in on the right side of the screen;
* During the second presentation, announced by the letter “B” on the screen, the PVS is on the left side and the SRC in on the right side of the screen.
* After the two presentation were completed the message “Vote” was shown for 5 seconds to allow viewers to write their score on the scoring sheet.

The above BTC organization was selected to minimise the influence of any possible un-uniformity in the display area.

## Meaning of the 11 grades scoring scale explained to the viewers

The 11 grade impairment scale was intended to allow to the viewers to express a judgement of the degradation (if any) between the “SRC” and the coded video clips (PVS).

A short training session was run for all the viewers so they could understand where and when to look at the screen and at the scoring sheet, and when to express their opinion.

The viewers were explained they had to carefully look at the video clips shown immediately after the message “A” and “B” was on the screen, to notice if they were able to see any difference to the video clip shown on the two half of the screen, being aware that the first presentation was with the SRC on the left side while during the second presentation the SRC was on the right part of the screen.

The following guidance was given about the numerical scoring:

* Should they not be able to see any difference between the source and the video clip “A” they had to write the number “10” in the box numbered with the same number of the message “Vote-N” they just saw on the screen.
* In the case that even a very small impairment was visible, they had to score 9, if the difference was just in one area of the image, or 8 if the difference was noted in many areas of the screen.
* 7 and 6 had to be used if the impairment were visible in a clear way.
* 5 and 4 were for when impairments were evident at first sight and easily.
* 3 and 2 were for when impairments were really annoying.
* 1 and 0 were for when the image was severely corrupted, in some area or everywhere.

## Test results

This paragraph shows the tables and the graphs of the subjective test conducted in Rome.

## Category 1 results

## Category 3a results

## Evidence for Category 1

## Evidence for Category 2

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | **ANCHOR** | | | |  | | | **P31** | | | | | | | |  | | **P32** | | | | | |  | | **P33** | | | | | |  | | **P34** | | | | | |
|  |  | ***MOS-CI*** | | ***MOS*** | ***MOS+CI*** |  | | ***MOS-CI*** | | | ***MOS*** | | | ***MOS+CI*** | | |  | | ***MOS-CI*** | | ***MOS*** | | ***MOS+CI*** | |  | | ***MOS-CI*** | | ***MOS*** | | ***MOS+CI*** | |  | | ***MOS-CI*** | | ***MOS*** | | ***MOS+CI*** | |
| S02 - Market | *Rate 1* | | 7,81 | 8,00 | 8,19 | |  | | | 9,10 | | | 9,20 | | | 9,30 | |  | | 8,30 | | 8,60 | | 8,60 | |  | | 8,19 | | 8,40 | | 8,61 | |  | | 8,68 | | 8,93 | | 9,19 | |
| *Rate 2* | | 7,30 | 7,53 | 7,76 | |  | | | 6,73 | | | 6,93 | | | 7,14 | |  | | 7,42 | | 7,73 | | 8,04 | |  | | 7,18 | | 7,47 | | 7,75 | |  | | 7,02 | | 7,40 | | 7,78 | |
| *Rate 3* | | 5,41 | 5,60 | 5,79 | |  | | | 6,03 | | | 6,40 | | | 6,77 | |  | | 5,83 | | 6,20 | | 6,57 | |  | | 5,88 | | 6,13 | | 6,38 | |  | | 5,25 | | 5,60 | | 5,95 | |
| *Rate 4* | | 4,32 | 4,73 | 5,14 | |  | | | 4,99 | | | 5,33 | | | 5,67 | |  | | 3,36 | | 3,73 | | 4,11 | |  | | 5,03 | | 5,20 | | 5,37 | |  | | 4,72 | | 5,00 | | 5,28 | |
| S05 - Show girl | *Rate 1* | | 8,00 | 8,20 | 8,40 | |  | | | 7,20 | | | 7,47 | | | 7,73 | |  | | 8,28 | | 8,47 | | 8,65 | |  | | 7,99 | | 8,27 | | 8,54 | |  | | 7,56 | | 7,80 | | 8,04 | |
| *Rate 2* | | 7,35 | 7,53 | 7,72 | |  | | | 6,82 | | | 7,13 | | | 7,45 | |  | | 7,07 | | 7,33 | | 7,60 | |  | | 7,03 | | 7,20 | | 7,37 | |  | | 7,09 | | 7,33 | | 7,58 | |
| *Rate 3* | | 5,81 | 6,00 | 6,19 | |  | | | 5,18 | | | 5,53 | | | 5,89 | |  | | 5,77 | | 6,10 | | 6,43 | |  | | 5,75 | | 6,00 | | 6,25 | |  | | 6,53 | | 6,80 | | 7,07 | |
| *Rate 4* | | 4,91 | 5,33 | 5,76 | |  | | | 3,99 | | | 4,27 | | | 4,54 | |  | | 3,84 | | 4,20 | | 4,56 | |  | | 4,57 | | 4,93 | | 5,30 | |  | | 4,76 | | 4,93 | | 5,11 | |
| S07 - Warm night | *Rate 1* | | 7,73 | 7,93 | 8,14 | |  | | | 7,28 | | | 7,53 | | | 7,78 | |  | | 7,71 | | 7,93 | | 8,16 | |  | | 7,49 | | 7,73 | | 7,98 | |  | | 7,18 | | 7,47 | | 7,75 | |
| *Rate 2* | | 7,09 | 7,27 | 7,44 | |  | | | 7,32 | | | 7,53 | | | 7,74 | |  | | 6,85 | | 7,13 | | 7,42 | |  | | 6,62 | | 6,87 | | 7,12 | |  | | 6,89 | | 7,07 | | 7,24 | |
| *Rate 3* | | 4,93 | 5,33 | 5,74 | |  | | | 5,49 | | | 5,80 | | | 6,11 | |  | | 5,27 | | 5,53 | | 5,80 | |  | | 5,49 | | 5,73 | | 5,98 | |  | | 5,37 | | 5,80 | | 6,23 | |
| *Rate 4* | | 3,96 | 4,20 | 4,44 | |  | | | 3,46 | | | 3,73 | | | 4,01 | |  | | 3,20 | | 3,53 | | 3,86 | |  | | 2,77 | | 3,07 | | 3,36 | |  | | 3,39 | | 3,60 | | 3,81 | |
| S08 - Baloon | *Rate 1* | | 7,95 | 8,07 | 8,18 | |  | | | 7,73 | | | 8,00 | | | 8,27 | |  | | 7,32 | | 7,53 | | 7,74 | |  | | 8,26 | | 8,47 | | 8,68 | |  | | 8,21 | | 8,40 | | 8,59 | |
| *Rate 2* | | 6,06 | 6,40 | 6,74 | |  | | | 6,73 | | | 6,93 | | | 7,14 | |  | | 5,87 | | 6,13 | | 6,40 | |  | | 6,39 | | 6,67 | | 6,95 | |  | | 6,70 | | 7,00 | | 7,30 | |
| *Rate 3* | | 5,46 | 5,73 | 6,01 | |  | | | 5,32 | | | 5,53 | | | 5,74 | |  | | 4,27 | | 4,53 | | 4,80 | |  | | 5,22 | | 5,53 | | 5,85 | |  | | 5,32 | | 5,60 | | 5,88 | |
| *Rate 4* | | 3,67 | 3,93 | 4,19 | |  | | | 3,40 | | | 3,67 | | | 3,93 | |  | | 3,20 | | 3,53 | | 3,86 | |  | | 3,83 | | 4,20 | | 4,57 | |  | | 4,15 | | 4,47 | | 4,78 | |
|  |  | |  |  |  | |  | | |  | | |  | | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |
| blue cells: | | | ***Proponent > Anchor*** | | | | | |  | | |  | | |  | | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |
| white cells: | | | *Proponent = Anchor* | | | |  | | |  | | |  | | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |
| red cells: | | | ***Proponent < Anchor*** | | | | | |  | | |  | | |  | | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |

## Conclusions

Form the evaluation performed in Rome it was possible to obtain two indication about the performances of the received Submissions:

1. There was clear evidence for many of the submissions, and most of the test conditions, of superiority in terms of visual quality at the same bit rates, compared to the existing technology.
2. It was possible to rank the submissions, on the basis of the visual quality provided by them, for both Category 1 and Category 3a.