|  |  |
| --- | --- |
| **Joint Collaborative Team on Video Coding (JCT-VC)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  23rd Meeting: San Diego, USA, 19–26 February 2016 | Document: JCTVC-W0090 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **HDR CE3: Results of subjective evaluations conducted with the DSIS method** | | |
| *Status:* | Input Document | | |
| *Purpose:* | Information | | |
| *Author(s) or Contact(s):* | Name(s) Martin Rerabek Philippe Hanhart Touradj Ebrahimi | Email: | [martin.rerabek@epfl.ch](mailto:martin.rerabek@epfl.ch) [philippe.hanhart@epfl.ch](mailto:philippe.hanhart@epfl.ch) |
| *Source:* | Ecole Polytechnique Fédérale de Lausanne (EPFL) | | |

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

# Abstract

This contribution reports the details and results of a second subjective quality evaluation conducted at EPFL to assess responses to the Call for Evidence (CfE) for HDR and WCG Video Coding [1]. Unlike our previous evaluations [2], we used the DSIS method for these assessments instead of a paired comparison to an Anchor. The same test material as in our previous evaluations was used in this subjective assessment. Results show similar trends as in [2], but less statistically significant differences are observed because of the lower discrimination power of DSIS approach when compared to paired comparison, as well as the color differences induced by the viewing angle dependency of the Sim2 monitor. These results can also be used to evaluate the performance of objective quality metrics.

# Introduction

To evaluate responses to the CfE for HDR and WCG Video Coding [1], we used a partial paired comparison method to have a direct answer to the following question: can a proponent achieve better visual quality when compared to an Anchor at a similar bit rate? However, the drawback of this method is that a direct comparison of different proponents cannot be made in a reliable way. For this purpose, a full paired comparison would be necessary, which requires tremendous efforts for the subjective evaluation. Another drawback was that the paired comparison results could not be used to measure the correlation between objective quality metrics and perceived visual quality. Thus, we have performed a second subjective quality assessment on the CfE material, but using DSIS methododology to obtain MOS values for all test stimuli. To be able to compare these new results with our previous results [2], as little changes as possible were made to the evaluation methodology and viewing environments.

# Subjective assessment

## Dataset

The same dataset as in [2] was used for subjective evaluation tests and consisted of five HD resolution HDR video sequences, namely, Market3, AutoWelding, ShowGirl2, WarmNight, and BalloonFestival. Similarly to [2], 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. The same cropping window as in [2] was used for each video sequence. However, unlike [2], the video sequences were played at their native frame rate, whereas they were all played at 24 fps in [2] due to the Pulsar’s fixed frame rate. In particular, sequence Market3 was played at 50 fps and all of its 400 frames were shown, whereas only the first 240 frames were shown in [2]. For the other sequences, the same frames as in [2] were selected. 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 | 0-399 |
| 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 8 bit AVI files, in the Sim2 packed format.

The side-by-side video sequences were generated using the HDRMontage tool from the HDRTools software [3]. Then, the conversion was made for the Sim2 display using the Sim2Convert tool provided in the HDRTools software [3].

## Test environment

The test was performed in the same environment as in [2]. In particular, the experiments were conducted at the Multimedia Signal Processing Group (MMSPG) test laboratory at EPFL, 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.

However, the test was performed on the full HD (1920 × 1080 pixels) 47'' Sim2 HDR47E S 4K HDR monitor. Similarly to [2], three subjects assessed the displayed test video content simultaneously in every session. They were seated in an arc configuration, at a constant distance of about 3.2 times the picture height (measured from the middle of the screen), as suggested in recommendation ITU-R BT.2022 [5].

## Methodology

The DSIS Variant I method with a five-grade impairment scale (*Very annoying*, *Annoying*, *Slightly annoying*, *Perceptible, but not annoying*, and *Imperceptible*) [4] was selected. 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 reference (unimpaired) video sequence. The other was the test video sequence, which is a reconstructed version of the reference.

To reduce the effect of position of video sequences on the screen, the participants were divided into two groups: the left video sequence was always the reference video sequence for the first group, whereas the right video sequence was always the reference video sequence for the second group. After the presentation of each pair of video sequences, a six-second voting time followed. Subjects were asked to rate the impairments of the test video sequence in relation to the reference video sequence.

## Test planning

Before experiments, oral instructions were provided to explain the evaluation task and a consent form was handed to subjects for signature. A training session was organized to allow subjects to familiarize with the assessment procedure. Similarly to [2], the same contents were used in the training session as in the test session to highlight the areas where distortions can be visible. Five training samples were manually selected by expert viewers, one for each level of the impairment scale and a different content for each sample. The samples were presented in the following order: *Imperceptible* (Market)*,* *Very annoying* (AutoWelding), *Annoying* (WarmNight), *Slightly annoying* (ShowGirl), and *Perceptible, but not annoying* (BalloonFestival). 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 4 test sessions. Each test session was composed of 50 basic test cells (BTC), corresponding to approximately 14 minutes each. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each group of subjects, whereas the same content was never shown consecutively. The test material was randomly distributed over the four test sessions.

Each subject took part in exactly two sessions. One dummy pair, whose score was not included in the results, was included at the beginning of the each session to stabilize the subjects' ratings. Between the sessions, the subjects took a 5-minute break.

A total of 30 naïve subjects (3 females and 27 males) took part in experiments, leading to a total of 15 ratings per test sample. Subjects were between 20 and 26 years old with an average and median of 22.9 and 23 years of age, respectively. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

# Statistical analysis

To detect and remove subjects whose scores appear to deviate strongly from the other scores in a session, outlier detection was performed. The boxplot inspired outlier detection technique proposed in [6] was used. In this study, no outlier subjects were detected. Then, the MOSs were computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% CIs, assuming a Student’s *t*-distribution of the scores.

To understand whether the difference between two MOS values corresponding to two different compression algorithms is statistically significant, a multiple comparison significance procedure was applied to the data [7]. Particularly, for each bit rate and content, a one-way ANOVA test was conducted to compare all compression algorithms pairwise to understand whether the differences of their means were statistically significant [7].

# Results and discussions

Figure 1 shows the resulting MOS/CI plots for different contents. As it can be observed, the CIs of the different proponents overlap in most cases, meaning that there are few cases where there is a statistically significant difference in visual quality. Nevertheless, improvements can still be observed, especially for proponents P11 and P22. Because the improvements over the Anchor are rather limited in many cases, they are harder to distinguish with an indirect comparison, e.g., DSIS, than with a direct comparison, e.g., paired comparison.

Surprisingly, the MOS for content ShowGirl2 are all below 4, except for Proponent P22 at R1. We believe that the scores for this content are lower because of the color differences induced by the viewing angle dependency of the Sim2 monitor and because subjects are more sensitive to color differences in human skin than for other regions. These results show that a simultaneous side-by-side presentation on Sim2 might not be suitable and that a DSIS temporal presentation may be considered as an alternative. However, the temporal presentation relies more on short-term memory and has a lower discrimination power than simultaneous presentation. The Variant II with repetition of each stimulus could be considered to compensate this effect.

|  |  |
| --- | --- |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsBitrate:Market3.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsBitrate:AutoWelding.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsBitrate:ShowGirl2.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsBitrate:WarmNight.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsBitrate:BalloonFestival.png | |

Figure 1 – Subjective results (MOS and CI). For content ShowGirl2, dummy values are used for one proponent (BBC), as the decoded material was not provided for this sequence.

Figure 2 shows the results comparing all possible pairs, for each bit rate separately. These results confirm that there are few cases with significant visual differences. In particular, Proponents P11 and P22 show significant improvements, especially at lower bit rates, whereas Proponent P14 is outperformed by most proponents on some contents.

|  |  |
| --- | --- |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MTPC:R1.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MTPC:R2.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MTPC:R3.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MTPC:R4.png |

Figure 2 – Results of the multiple comparison tests for different test conditions, i.e., combination of algorithm and bit rate (R1 to R4). In each plot, the color of each square shows the number of times (i.e., for how many contents) the MOS corresponding to condition A is statistically significantly better than the MOS corresponding to condition B.

# Conclusion

In this contribution, the details and results of the subjective evaluation conducted at EPFL using the DSIS method to evaluate the responses to the CfE for HDR and WCG Video Coding have been reported. Results suggest that a sequential instead of a simultaneous presentation could be considered because of the color differences induced by viewing angle dependency of Sim2 monitor. The subjective results can be used to evaluate performance of objective quality metrics.

# Patent rights declaration(s)

**EPFL does not have any current or pending patent rights relating to the technology described in this contribution.**

**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, “Results of HDR CfE subjective evaluations conducted at EPFL,” Doc. M36728, Warsaw, Poland, July 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.
6. F. De Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi, “Towards high efficiency video coding: Subjective evaluation of potential coding technologies,” Journal of Visual Communication and Image Representation, vol. 22, no. 8, pp. 734–748, November 2011.
7. G. W. Snedecor and W. G. Cochran, Statistical Methods, Iowa State University Press, 1989.