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/M37319
October 2015, Geneva, Switzerland

Source [1]University of British Columbia (UBC) & [2]TELUS Communications Inc.
Status Input Document
Title **Evaluation of Scalable versus Single Layer Compression on Consumer HDR Displays**
Author [1]Ronan Boitard, [1]Maryam Azimi, [1, 2]Mahsa T. Pouzarad, [1]Panos Nasiopoulos

Abstract

In this work, we evaluate the compression of HDR content using either a single layer or scalable approach. Content were evaluated on a 10-bit per channel, P3 display with a peak luminance of 1,000 nits. Evaluated content was graded on a display with higher peak luminance than the one used in this work. Simulations were compliant with the anchors defined in the HDR Call for Evidence (CfE) [1] and the subjective evaluations followed the guidelines of the CfE. Results point towards a higher efficiency of the single layer compression scheme, however the amount of increased efficiency was really hard to judge since the CfE suggested tests conditions were not suited for the task in hand. More specifically, for three out of five sequences, the single layer tends to outperform the scalable approach. For the other two sequences, the difference in quality between the lower and higher bit rates is not significant enough to reach any conclusion. Furthermore, comparison with tests performed on the SIM2 shows that using a display with lower peak luminance can smooth out compression artifacts. We plan to address the above challenges in future contributions

1 Introduction

Distributing HDR and SDR content to the end-users has been the focus of many discussions over the past months. Several distribution schemes have been proposed to address both types [2]. These different distribution techniques can be classified as follows:

1. Compression of HDR content; color and tone mapping are performed at the decoding stage to generate SDR (Figure 1-a): additional metadata can be derived to improve the quality of the color and tone mapping (Figure 2-b),
2. Compression of SDR content with inverse color and inverse tone mapping at the decoding stage to generate HDR (Figure 1-b); additional metadata can be derived to improve the quality of the inverse color and inverse tone mapping (Figure 2-a),
3. Compression of HDR and SDR separately (Figure 3-a), denoted as simulcast (two separate single layers),
4. Joint compression of HDR/SDR content (Figure 3-b), denoted as scalable.

In this informative document, we propose to compare the efficiency of scheme 3 (HDR Single layer) and 4 (Scalable).

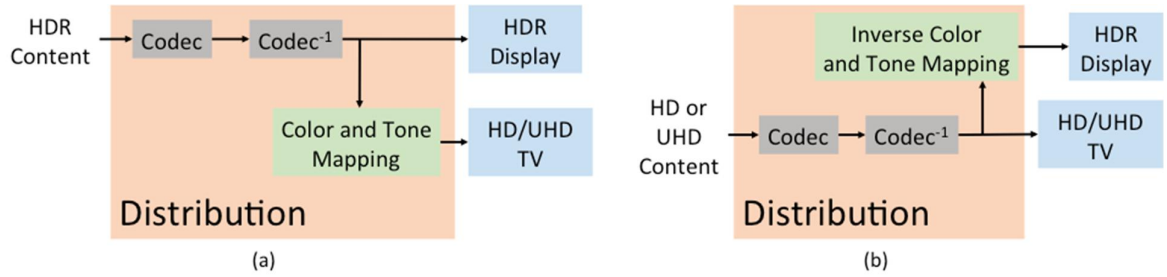


Figure 1: Distribution of HDR content and derivation of SDR content at the display stage (a) and distribution of SDR content and derivation of HDR content at the display stage (b).

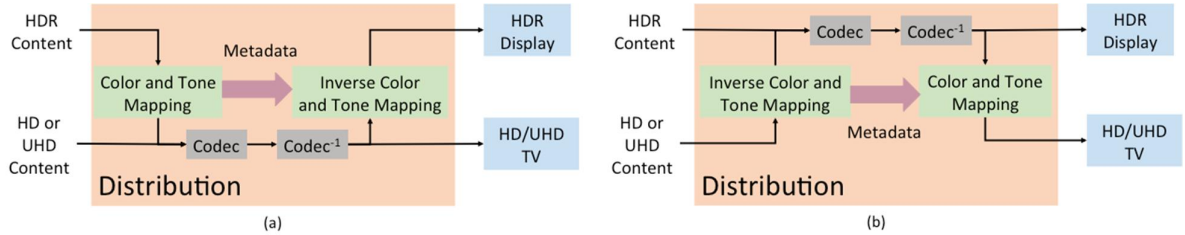


Figure 2: Use of metadata for reconstructing distributed content in several versions.

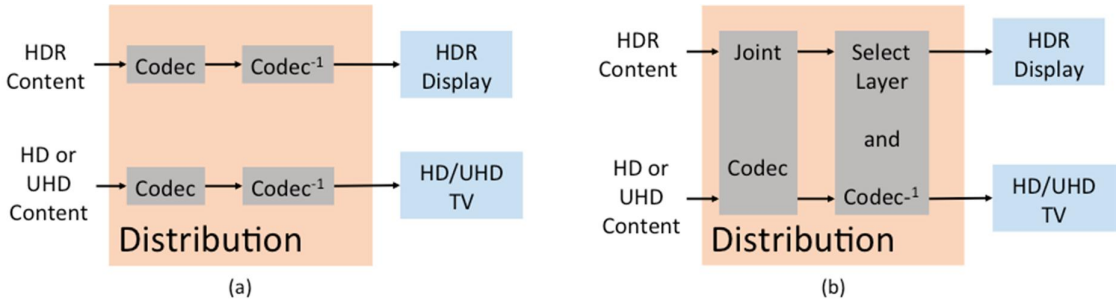


Figure 3: Simulcast (a) and Scalable (b) distribution when several inputs of the same content are available.

2 Content

To perform our tests, we used 5 sequences considered in the Call for Evidence (CfE) [1]. Three different versions of these sequences were considered: HDR10, SDR_A10 and SDR_C10. The HDR10 sequences correspond to the original HDR sequence of images (OpenEXR or Tiff 16 bpc) that have been perceptually encoded using the SMPTE ST 2084 [3] with non-constant luminance and quantized on 10 bits after chroma subsampling to 4:2:0 (see Figure 4).

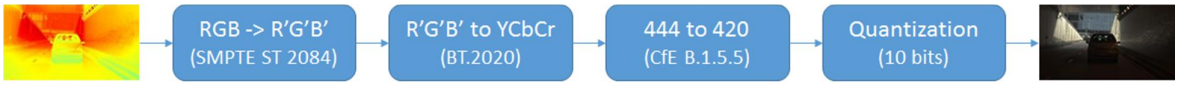


Figure 4: Preprocessing of original HDR content to obtain HDR10 source.

SDR_A10 and SDR_C10 correspond to the SDR sequences presented in the CfE [1]. Table 1 summarizes the sequences used and their correspondent acronyms in [1].

Note that the sequence Market3 was voluntarily removed because its frame rate (50 fps) was too high for our test setup (see 3.2 for details).

Sequence	HDR10	SDR_A10 (Class – Seq.)	SDR_C10 (Class – Seq.)
FireEater2	Generated	AA – SA00	AA – SC00
Tibul2	Generated	AA – SA01	AA – SC01
AutoWelding	Generated	N/A	AA – SC03
BikeSparklers	Generated	N/A	AA – SC04
BalloonFestival	Generated	AA – SA08	AA – SC08

Table 1: Summary of source sequences used for the tests

3 HDR Test Scenario

4.1Conditions

Three different configurations, using different QPs, have been used to obtain the compressed content:

1. SM10: HDR10 sources compressed using HEVC (HM 16.6),
2. SCC10_L1: HDR10 and SDR_C10 sources using HEVC (SHM 0.8),
3. SCA10_L1: HDR10 and SDR_A10 sources using HEVC (SHM 0.8),

The different compression points are summarized in Table 2. Note that FireEater2 SCC10_L1 QP = 20 does not corresponds to the anchor (QP = 18) since the SHM 0.8 was crashing whenever this QP was used.

Sequence name	SM10 - QPs	SCC10_L1 - QPs	SCA10_L1 - QPs
FireEater2	20,23,26,29	20,23,26,31	20,23,26,29
Tibul2	19,24,29,34	19,24,29,34	19,24,29,34
AutoWelding	21,25,29,33	21,25,29,33	N/A
BikeSparklers	23,25,29,33	23,25,31,35	N/A
BalloonFestival	18,22,26,30	18,22,26,30	18,22,26,30

Table 2: Chosen QPs for each sequence and configuration.

4.1HDR Experimental Study

Using the decoded content, we conducted an experiment to assess the quality of the decoded sequences compared to the HDR10 source. The experiment was conducted on a Samsung SUHDTV UN65JS9500 series 9 of resolution 3840x2160 which is a 65” 10-bits commercial TV with a peak luminance of 1,000 nits and a P3 color gamut. The set was modified to accept HDR input via an HDMI input. The input signal is encoded to R’G’B’ using the SMPTE ST 2084 [3] and represented in a BT.2020 container.

We used a 12 bits per channel HD (1920x1080) signal at 30 frames per seconds (graphic cards limitations). The Scratch player [4] was used to reproduce up to 10 bits. Since the resolution cannot encompass two HD videos, we cropped each sequence along the horizontal axis with the

horizontal coordinates indicated in Table 3. The corresponding low dynamic range images using the same cropping are provided in Annex 1.

Sequence name	Cropped area
FireEater2	550 - 1497
Tibul2	800 - 1747
AutoWelding	375 - 1322
BikeSparklers	550 - 1497
BalloonFestival	1 - 948

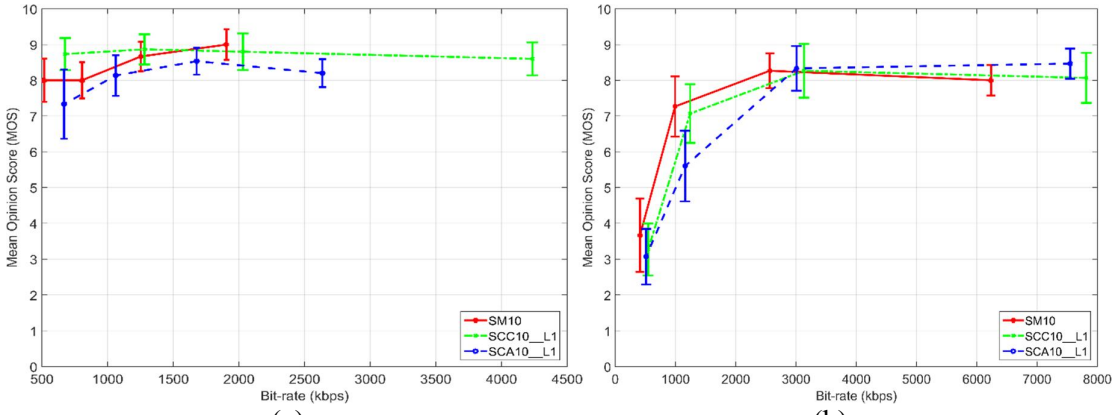
Table 3: Cropped area per sequences.

Note that the peak luminance of content is higher than that of the TV (4,000 nits vs. 1,000 nits), thus we scaled down (dividing by 4) each color channel in the linear domain. This scaling was chosen so as to not clip any information and preserve highlights. It could be argued that higher quality reproduction may have been achieved using a tone curve that adapts to the content. However such an approach might affect compression artifacts differently depending on the sequence. Displaying directly HDR content (not performing scaling) resulted in severe clipping and, thus, was not included in our tests.

The subjects of the experiment were asked to evaluate the quality of decoded content compared to original. A scale from 1 to 10 was used to assess the quality. A training session was organized before the actual experiment to describe compression artifacts. The evaluation was composed of 52 ratings (5 sequences × 4 QPs × 2 pipelines + 3 sequences × 4 QPs)that were randomly ordered so that two same sequences would not be shown following each other. Subjects were given time between two stimulus to vote.

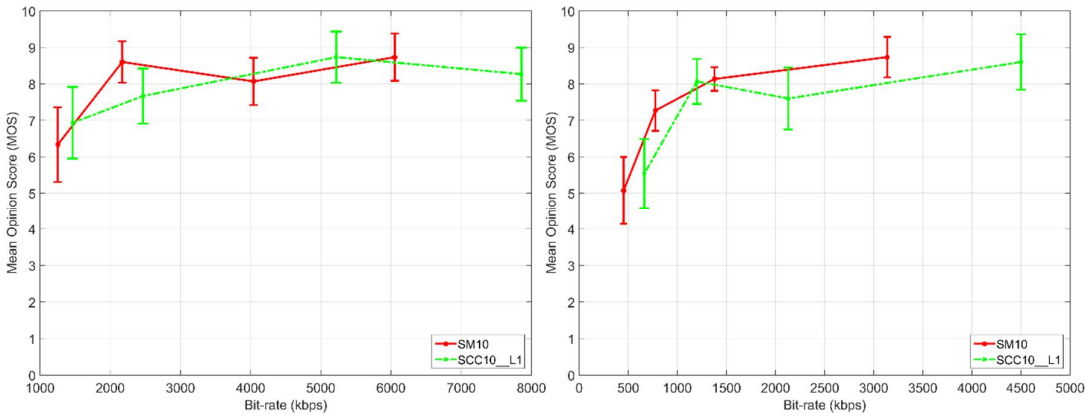
4 Results

20 subjects (with 5 outliers) took part in this experiment and all were screened for color blindness and visual acuity. The results are provided in Figures 5 to 7. We observe that for the case of Tibul2, BalloonFestival, and AutoWelding, the single layer tends to outperform the scalable approach. However, because the bit rates for the two configurations are different, extrapolating the amount of increased efficiency is not straightforward.
Note that for FireEater2 and BikeSparklers, the difference in quality between the lower and higher bit rates is not significant enough to reach any conclusion. This is due to the fact that the QPs used were not high enough to introduce any visible distortions for these two streams.



(a) (b)

Figure 5: Mean Opinion Score (MOS) for (a) FireEater2 and (b) Tibul2



(a) (b)

Figure 6: Mean Opinion Score (MOS) for (a) BikeSparklers and (b) AutoWelding

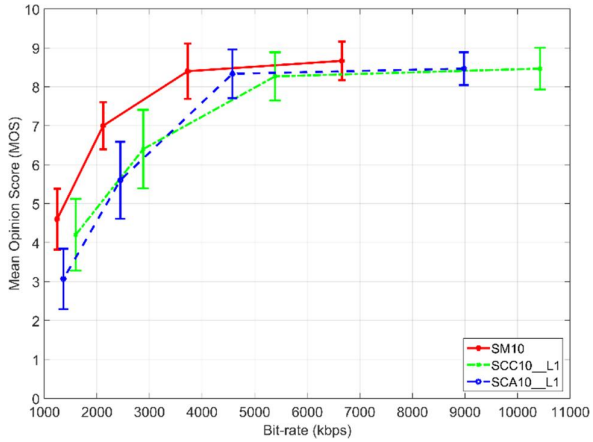


Figure 7: Mean Opinion Score (MOS) for BalloonFestival

4.1 Discussion

For some sequences, such as FireEater2, no difference in quality could be noticed independently of the bit-rates for both pipelines. Due to the limited time to perform the same evaluation on a different display, only 4 expert observers watched this sequence on both the SIM2 and the Samsung TV. They reported that visible degradation in quality could be observed between each QP on the SIM2 display while no such degradation was noticeable on the Samsung TV. Note that the original content was graded for the SIM2 and, thus, remapping was needed for displaying it on the lower peak luminance display. This may be the reason why compression artifacts were not as visible on the Samsung display.

Obviously these results are to be taken with caution, since they could greatly depend on the used scaling/tone-mapping. We believe that at this point more tests with different tone mapping approaches using different displays are necessary in order to identify a more general trend.

5 Conclusion and Recommendations

In this informative document, we tried to assess the difference in performance between scalable and single layer compression for HDR content. Reported results show that the current conditions as defined in the MPEG Call for Evidence do not allow evaluation of the difference between those two schemes with certainty. Based on our observations, we would recommend adding higher QPs (lower bit-rates) for most of the sequences. Furthermore, the methodology used seems to provide very high confidence intervals, reducing the viability of the results. An approach that could yield better results would be to obtain similar bit-rates for the different configurations and perform a pair-wise comparison (with or without force choice).

In addition, we would like to recommend that when content is graded on a display, it should always be tested for compression efficiency on this display and optionally be tested on different displays.

Finally, bear in mind that in a typical broadcast scenario, the display at the end user is different from the one on which the content was graded. Although evaluating the tone mapping is out of scope, evaluating the effect of compression efficiency on a variety of displays with different limitations is an important issue. Our tests have shown that when tone mapping is required, the compression artifacts seen on a high peak luminance display are less visible on a lower peak luminance display that scales the content.

6 References

[1] A. Luthra, E. Francois, and W. Husak, “Call for Evidence (CfE) for HDR and WCG Video Coding,” in ISO/IEC JTC1/SC29/WG11 MPEG2015/N15083, 2015.

[2] **In Press:** R. Boitard, M. T. Pouzarad, and P. Nasiopoulos "Demistifying High Dynamic Range," in IEEE Consumer Electronics Magazine, 2015.

[3] Society of Motion Picture & Television Engineers, "High Dynamic Range Electro- Optical Transfer Function of Mastering Reference Displays Table," in SMPTE ST 2084, 2014.

[4] Scratch Player: http://www.assimilateinc.com/.

7 Annex 1: Cropping