EBU

OPERATING EUROVISION AND EURORADIO

DATE
19th November 2014

Mr Patrick LeCallet and
Mr Philip Corriveau
VQEG, HDR co-chairs

BY EMAIL
patrick.lecallet@polytech.univ-nantes.fr
philip.j.corriveau@intel.com

REFERENCE
DTI/191114/yt

SUBJECT
EBU liaison letter on HDR activities: Development of methodology for assessing the quality of HDR video

Dear Mr LeCallet and Mr Corriveau,

The European Broadcasting Union (EBU) notes that the aim of the VQEG HDR project is to develop methods for assessing the quality of HDR.

The EBU would like to be kept informed about further progress toward this goal and encourage VQEG to fulfil this goal within the next 6 months.
The EBU Strategic Programme on BeyondHD (SP BHD) strongly encourages VQEG-HDR to develop an appropriate methodology for Higher Dynamic Range images.

To inform the debate and as a basis for VQEG`s work, you may like to know that the EBU`s Strategic Programme on BeyondHD (SP BHD), in conjunction with the EPFL[1] and Dolby Laboratories Inc., has conducted tests with a variety in testing methodologies. The conclusion we reached is that the methods we employed would allow consistent evaluation of Higher Dynamic Range (HDR) images. However, further investigations are needed to fully develop a suitable test methodology, e.g. covering such issues as how to adjust the monitor black level in different viewing environments, guidelines for grading the content and for comfortable viewing.

Yours sincerely,

Giorgio Dimino, RAI
(Chairman of SP BeyondHD)

Dagmar Driesnack, IRT
(Vice-Chair of SP BeyondHD)

Cc: Yvonne Thomas (EBU Coordinator of SP BeyondHD)
Hans Hoffmann, EBU (coordinator of internal UHDTV activities)
Margaret Pinson, U.S. Dept. of Commerce, NTIA/ITS, Boulder, Colorado U.S.A.
Arthur Webster, U.S. Dept. of Commerce, NTIA/ITS, Boulder, Colorado U.S.A.

[1] École polytechnique fédérale de Lausanne

EBU

OPERATING EUROVISION AND EURORADIO

Annexe

Proposal for an evaluation method for higher dynamic range content

1. Test outline

The tests did not focus on evaluating the added value of EIDR, but on identifying an appropriate methodology for EIDR evaluation because no standardized methodology to evaluate EIDR currently exists. It is important that the chosen methodology can identify the effect of Extended Image Dynamic Range in isolation from screen brightness, colour gamut etc.

Before taking part, the subjects' colour vision was checked using standard Ishihara and Snellen vision tests. Those subjects that did not pass the colour vision check (e.g., colour-blind) were not allowed to participate in the evaluation. If subjects normally wear glasses or contact lenses in their daily life, they were advised to wear them during the evaluation.

Following the colour vision test a training session was given. This consisted of oral instructions to explain the task and allow the subjects to familiarize themselves with the assessment procedure. This was followed by two video sequences that demonstrated the procedure, using different versions: 4000 nits, 1000 nits, 400 nits, and 100 nits.

At the EBU test a time sequential playback of the test sequences has been chosen and presented to expert viewers.

The EPFL test conducted the evaluation with naïve viewers in a Side by Side presentation on the same screen.

Only the relevant test set-up and the scoring (see point 3) were explained to the subjects during the training sessions.

The tests were arranged such that five subjects at the EBU and four subjects at the EPFL were evaluating the EIDR material during each test session.

Arrangement,

- Two subjects sat at a viewing distance of 1,5 m (equal 3H)
- Three subjects at the EBU and two subjects at the EPFL sat at a viewing distance of 2,7 m (equal the average domestic viewing distance in the UK)[2].

The monitor has been adjusted so that the eye height of subjects was at approximately horizontal middle of the screen. The subjects were seated in checkerboard style, so they did not obstruct each other's view of the display.

2. Test setup

For the evaluations Dolby Laboratories Inc. provided one of their 42" Pulsar monitors with a peak brightness of 4000 nits, P3 colour gamut and PQ EOTF. For the evaluation it has been decided to use Rec. 709 colorimetry in the test sequences in different versions regarding peak brightness:

[2] Tanton, N.E., "*Results of a survey on television viewing distance".* R&D White Paper No. 90, British Broadcasting Corporation, London, 2004

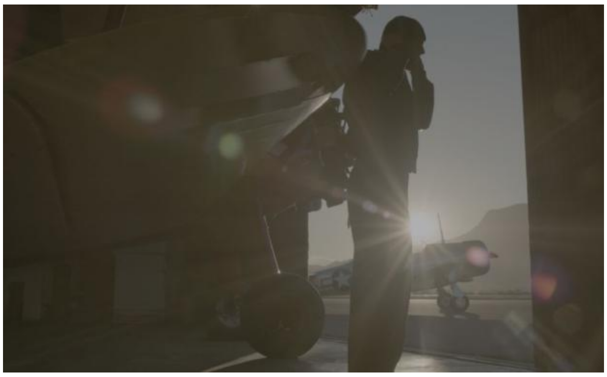
1) Manually graded at 4,000 nits (reference), displayed at 4,000 nits
2) Above *Content Mapped* to 1,000 nits, displayed so mean brightness was similar to reference.
3) Above, *Content Mapped* to 400 nits, displayed so mean brightness was similar to reference.
4) Above, *Content Mapped* to 100 nits, displayed so mean brightness was similar to reference.

It should be noted that the content mapping from 4000 nits to 1000, 400 and 100 nits may not be representative for content that would have been originated in this luminance.
The six test sequences (see figure 1) were presented in 1080p resolution for a duration of 20 s for both, Side by Side and Time sequential presentation.

The EBU set the illumination surrounding the display to 10 nits for most test groups and to 24 nits for one test group in order to get a wider feedback. The EPFL has set their surround to 20 nits and was thus in the same range of the EBU test backlight settings.

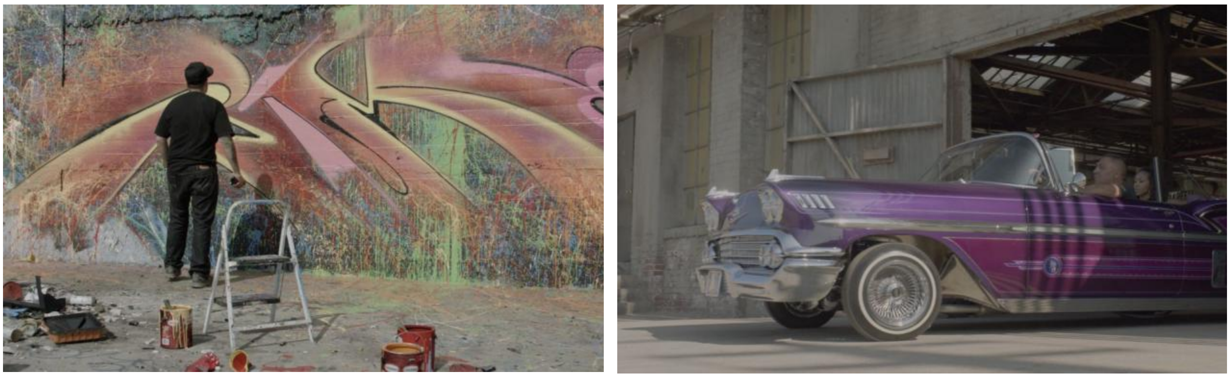
Clip 1: Flower Field

Clip 2: Plane



Clip 3: Sun

Clip 4: Sparkles



Clip 5: Art3

Clip 6: Car garage

Figure 1: Screenshots of the six sequences under test

3. Scoring - images of the scoring sheets and number of participants for each version including EPFL

The tests conducted by the EBU and EPFL have a forced choice with a horizontal preference scale, as shown in figure 2 and 3. In the EBU test “Left” was replaced with “A” and “Right” with “B”.

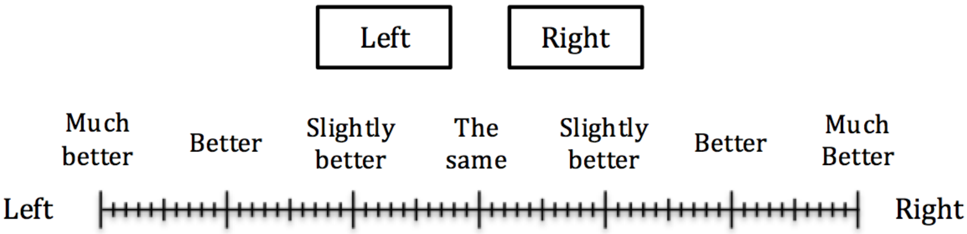


Figure 2: Scoring scale used in EPFL subjective evaluations

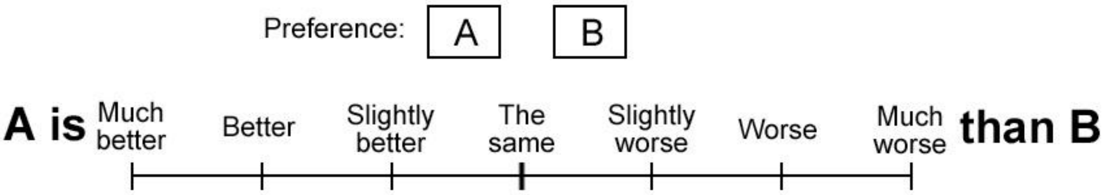


Figure 3: Scoring scale used in EBU subjective evaluations

Subjects were asked to rate the overall quality of a series of video clip pairs that differed in one parameter (EIDR). Differences that could be noticed between clip pairs included, but were not limited to: the overall image colour, the quality of the reproduction of skin tones, the details of shadows in the scene, the contrast and the details of highlights or other light sources appearing in the scene.

The complete evaluation lasted approx. 50 minutes for the EBU test in Time sequential presentation and 30 minutes for the EPFL test in Side by Side presentation.
For each trial the subjects saw 2 variations of the same source video clip (A & B sequential, or left & right simultaneous). The order of the video clips across trials and groups was randomized.
For each vote, each clip were shown twice in an A-B-A-B time sequential sequence at the EBU, as shown in figure 4, with

T1 = 20 s Test sequence A
T2 = 3s Mid-grey
T3 = 20 s Test sequence B
T4 = 5 s Mid-grey

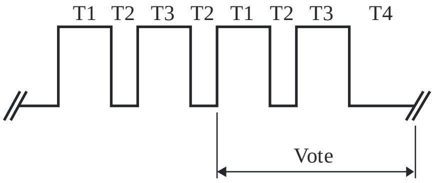


Figure 4: Double stimulus continuous scale method – Trial structure[3]

At the EPFL for each vote, each clip was shown twice in Side by Side before voting.
While voting subjects had first to identify which of the two video clips (A & B or left & right) they prefer by selecting the “A or B” or “Left or right” box at the top of the scoring scale on the scoring paper. A second step for each vote included to indicate the tendency of how Left relates to Right, respectively A relates to B, on the continuous scale.

4. Conclusions

The EBU had submitted conclusions from the EBU/EPFL tests for discussion within the ITU RG-24. Results obtained from both evaluations were highly correlated which proves they offer a good degree of reliability and reproducibility in different premises and test environments. Analysis of the scores in both cases also show good confidence intervals for each point under test.
The voting results were analyzed. The preference score for all graded versions compared to the 4000 nits version (reference) was computed . By normalizing the scoring scale, the preference probability was computed for all video versions compared to reference 4000 nits as illustrated in figures 5 and 6 below. Readers should note that, whilst indicative of the quality of the methodology used, the non-reference variants were automatically generated (and, therefore, are not representative of what a human colorist could achieve) and work still needs to be completed on defining how images graded to different luminance levels can be displayed at a single, defined average brightness. **Figures 5 and 6 should, therefore, not be used for drawing conclusions regarding the quality increase seen by using Extended Image Dynamic Range.**
When comparing the EBU Preference scale vs. the EPFL Preference scale the

[3] Please note that the correlation was computed on the 6 common sequences and 4 grades, resulting in a total of 24 test points.

- Pearson linear correlation coefficient is 0.9505[4]
- Spearman rank order correlation coefficient is 0.8913[5]

The horizontal preference scale can be used as an appropriate evaluation method. Compared to the forced choice method, which also proved to be a valid evaluation method, the preference scale shows a higher accuracy in the confidence intervals and is thus preferred.

The EBU has separately tested the DSCQS scale which also returned reliable results but were less consistent than the continuous preference scale as described in this document.
The results have also shown that a difference in EIDR is visually recognized independent of the viewing distance. However, a concrete quantification on the added value of EIDR level has not been evaluated and should be the subject of further tests.
Further tests need to be conducted in regard to the ambient light, as a lower ambient level was commented as too low and the highest as too bright.
For upcoming tests we also need to distinguish between high display brightness and an Extended Image Dynamic Range.
Readers of this document should also note that some viewer comments reported discomfort caused by the high brightness (4000 nits image) in subareas of the images.

[4] http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf

[5] http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

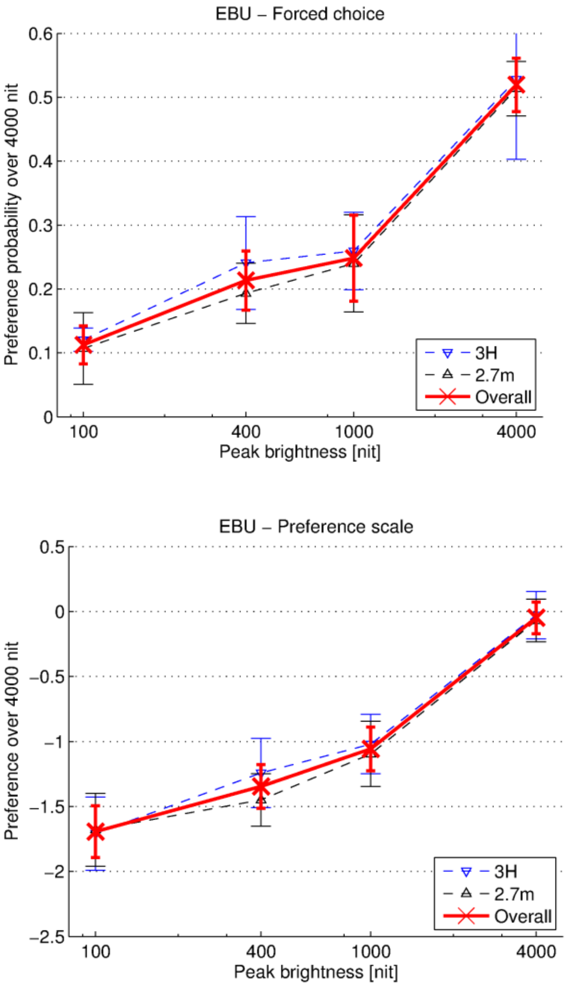


Figure 5: Forced choice (above) and preference (below) of various grading versions, 100 nits, 400 nits, and 1000 nits, compared to reference 4000 nits version computed from EBU evaluations results

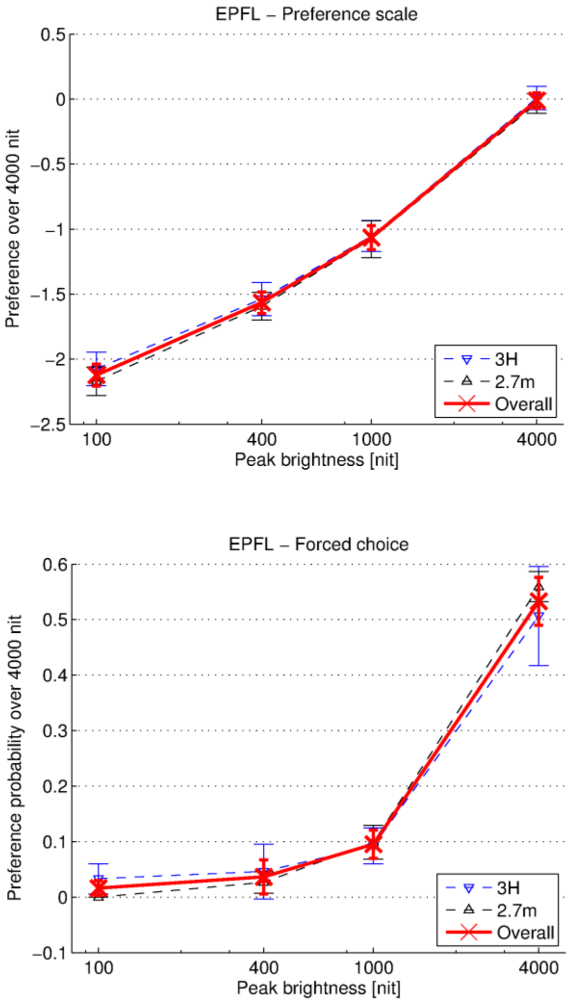


Figure 6: Forced choice (above) and preference (below) of various grading versions, 100 nits, 400 nits, and 1000 nits, compared to reference 4000 nits version computed from EPFL evaluations results