SUBJECTIVE QUALITY EVALUATION OF HIGH DYNAMIC RANGE VIDEO AND DISPLAY FOR FUTURE TV

Philippe Hanhart, Pavel Korshunov and Touradj Ebrahimi

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Yvonne Thomas and Hans Hoffmann

European Broadcast Union (EBU), Switzerland

ABSTRACT

The main objective of this paper is to verify test methodologies for assessment high dynamic range (HDR) video. To achieve this, a next generation HDR monitor by Dolby Laboratories was used to display professionally produced HDR content, including some scenes from recently released movies. Two complementary approaches for subjective assessment of HDR video were then designed and carried out at the EBU and EPFL premises. Results obtained from both evaluations were highly correlated which show they offer a good degree of reliability and reproducibility in different situations. Analysis of the scores in both cases also show good confidence intervals for each point under test. Finally, they could demonstrate that an increase in terms of quality of experience can be expected from the conventional level of 100 nits to HDR/high brightness at 4’000 nits, with intermediate improvements at 400 and 1’000 nits.

INTRODUCTION

Several technological revolutions have impacted the television industry over the last decades, such as the shifts from black & white to color and standard to high definition. Nevertheless, considerable improvements may still be achieved along several orthogonal axes, including resolution, color, frame rate, contrast, and brightness. Until recently, 3DTV was advertised as the future of television. However, due to lower picture quality and the need to wear cumbersome glasses, 3DTV has not fulfilled customer satisfaction yet. The momentum behind ultra high definition TV (UHDTV) is quickly building up, and in particular during the last couple of years, but some believe that UHDTV could risk to experience the same reaction as in 3DTV. With the recent advances in display technologies [1,2], high dynamic range (HDR) imaging has gained increased interest, and with that the concept of high dynamic range TV (HDRTV).

HDRTV allows rendering a wider range of luminance values to better represent details in both dark and bright areas, which is closer to what the human eye can perceive. An important question is what are the real impacts of these enhancements on viewers’ quality of experience. This paper attempts to open the door towards answering to this question, by proposing new quality evaluation methodologies for HDR video. In fact, efficient test methodologies are essential in this task. To achieve this, carefully selected video sequences at four different peak luminance levels were displayed either sequentially or side-by-side on a Dolby Research HDR RGB backlight dual modulation display (aka Pulsar), capable of the accurate and reliable reproduction of color and luminance. The black level was held constant, so that the luminance dynamic range was solely determined by the maximum luminance. The tested luminance levels reflect four levels of dynamic range that are typical for current and future consumer scenarios, given today's current displaying technologies and latest advances in HDR displays. Based on these, two alternative quality assessment methodologies were designed to seek for highly accurate and reliable measures of perceptual preferences. A first evaluation methodology, carried out at the EBU premises, relied on expert subjects and was carried out in a time sequential display mode, while a second was carried out at EPFL by naïve viewers in side-by-side. In the remainder of this paper, details on the design and implementation of each methodology are provided, and their results are presented and compared. Conclusions are drawn at the end of the paper.

ASSESSMENT METHODOLOGIES

The primary purpose of the work reported in this paper is to design assessment methodologies to evaluate the quality of HDR video sequences, and to verify their performance in terms of reliability and repeatability. However, as a by-product, preliminary results on the added value of HDR video sequences are also reported.

In both assessment methodologies reported in this paper, subjects' colour vision was checked using standard Ishihara and Snellen vision tests. Those subjects that did not pass the vision check (e.g., color blind) were not allowed to participate in the evaluations. If subjects wore glasses or contact lenses in their daily life, they were advised to wear them during tests. A training session was also organized during which oral instructions were provided to subjects explaining their task and at the same time to allow them to familiarize with the assessment procedure.

Evaluations carried out at the EBU relied on expert viewers, comparing video material specifically tailored for display at 100, 400, 1'000 and 4'000 nits. These four dynamic range levels were selected to be representative of key use cases, as opposed to being uniform perceptual distances, and constructed as follows:

1) 4'000 nits version: manually graded by professional colorists from the original captured video.

2) 1'000 nits version: tone-mapped from manually graded 4'000 nits version.

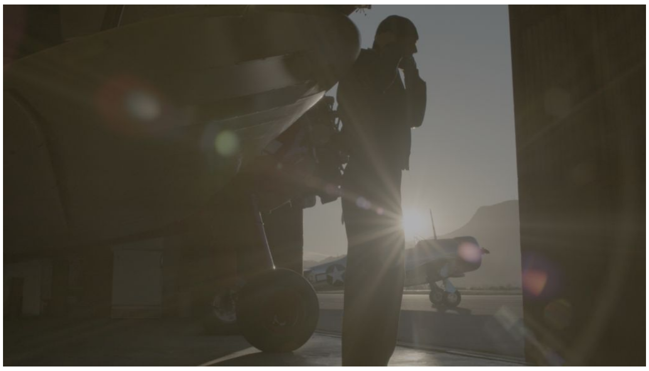
3) 400 nits version: tone-mapped from manually graded 4'000 nits version.

4) 100 nits version: tone-mapped from manually graded 4'000 nits version.

For tone-mapping, an automated proprietary tone-mapping algorithm was used. This algorithm was designed to preserve overall appearance to the input (graded) version. It was not intended to do enhancement.

Clip 1: Flower Field

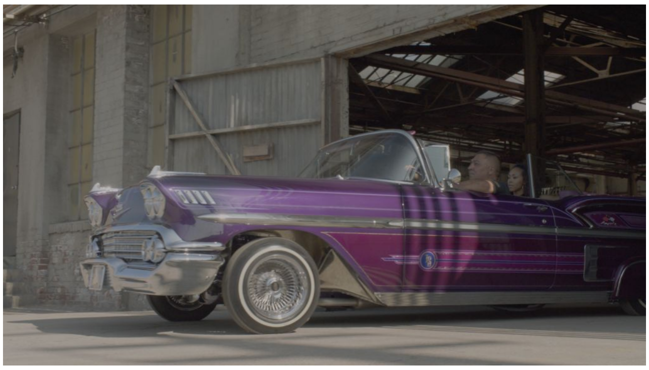
Clip 2: Plane



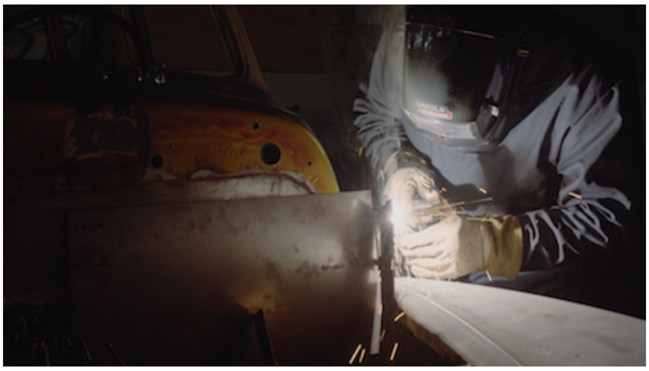
Clip 3: Sun

Clip 4: Sparklers

Clip 5: Art3



Clip 6: Car garage



Clip 7: Welding (EPFL test only)

Figure 1: Representative frames of the sequences. Clips 1 to 6 were used in both tests. One additional Hollywood movie clip used in the EPFL test is not included due to its Copyright.

A time sequential presentation with both a forced-choice preference, as well a horizontal preference scale, was conducted, using a 4’000 nits graded content as a hidden reference. This methodology was derived from the stimulus comparison methodology [3]. The forced-choice preference is a binary scale that directly identifies which condition is preferred, whereas the horizontal preference scale, which is a continuous version of the comparison scale [3], provides a finer comparison of the two conditions. Two viewing distances, of 3H (1.5m) and 2.7m, were tested. Evaluations carried out at the EPFL were similar to those at the EBU, but used naïve subjects and were performed with a side-by-side simultaneous presentation. These also employed a forced-choice response as well as a horizontal preference scale.

In the EBU tests, six test sequences (see figure 1) were presented in 1080p resolution and a length of 20 s in a time sequential presentation. In the EPFL tests, in addition to the above six sequences, two more sequences were also included, presented in 1080p resolution and a length of 20 s in side-by-side presentation.

At the EBU tests, the illumination surrounding the display was set to 10 nits for most test groups, and then to 24 nits for one test group in order to get a wider feedback. At the EPFL tests, the illumination surrounding the display was set to 20 nits and was thus in the same range as in the EBU tests backlight settings.

The scores in the EBU and EPFL tests included both a forced choice and horizontal preference scales, as shown in figure 2. In the EBU tests “Left” was replaces with “A” and “Right” with “B” because of the time sequential display mode used.

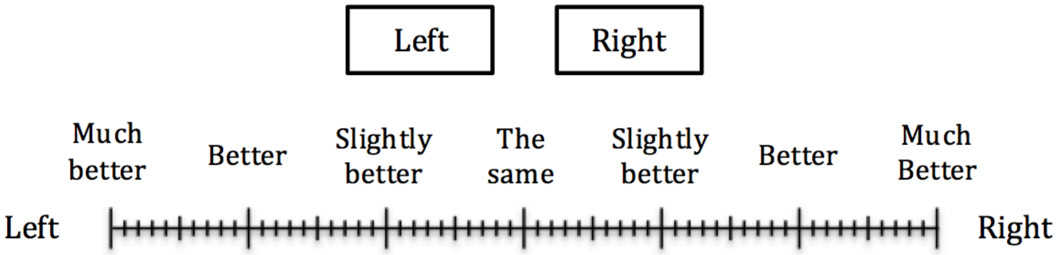


Figure 2: Scoring scales used in the tests

(in the EBU tests “Left” was replaced by “A” and “Right” by “B”).

Subjects were asked to rate the overall quality of pairs of displayed video sequences. To select a score, subjects were instructed to consider characteristics such as color rendition, quality of the reproduction of skin tones, details of shadows in the scene, contrast and the details of highlights, presentation of light sources appearing in the scene, etc.

Each evaluation session lasted approx. 50 minutes for the EBU tests in time sequential mode and 15 minutes in the EPFL tests in side-by-side presentation.

For each trial subjects saw 2 variations of a same source video clip (A & B sequential, or left & right side-by-side). The order of the video clips across trials and groups was randomized. For each vote, each clip was shown twice in an A-B-A-B time sequential mode at the EBU tests, as shown in figure 3, with

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

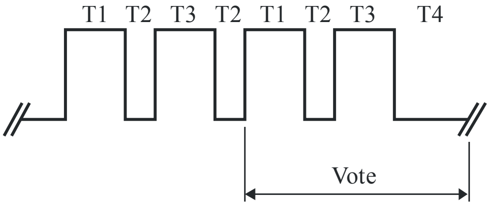


Figure 3: Time sequential mode presentation at the EBU tests

At the EPFL tests, each video sequence was shown only once in side-by-side before voting.

RESULTS

Figure 4 shows the overall results obtained at the EBU and the EPFL tests for forced choice scores, with their respective confidence intervals. It can be seen that the forced choice preferences increase with the increase in the peak brightness from 100 nits to 4’000 nits, to reach 0.5, which corresponds to the preference of 4’000 nits sequences to themselves that should be theoretically random (50%). These results show that there is a significant preference towards 4’000 nits displayed content when compared to other alternatives considered in the tests.

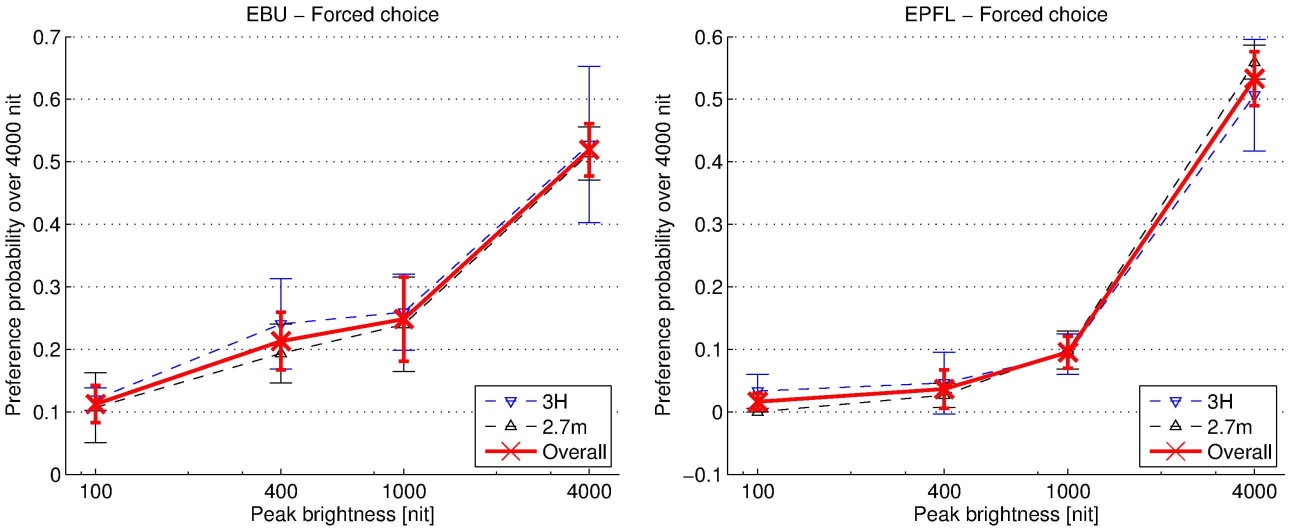


Figure 4: The EBU and EPFL results for forced choice scores

Figure 5 reports the overall results obtained at the EBU and the EPFL tests for preference scores to 4’000 nits reference, and their respective confidence intervals. Results confirm the same trend as in the forced choice case, but also better quantify the steps in terms of preference for intermediate peak brightness contents. In particular, the tighter confidence intervals in the EPFL tests show that a side-by-side display provides not only more reliable scores but also better quantifies the differences between 400 nits and 1’000 nits in terms of preference when compared to 4’000 nits reference.

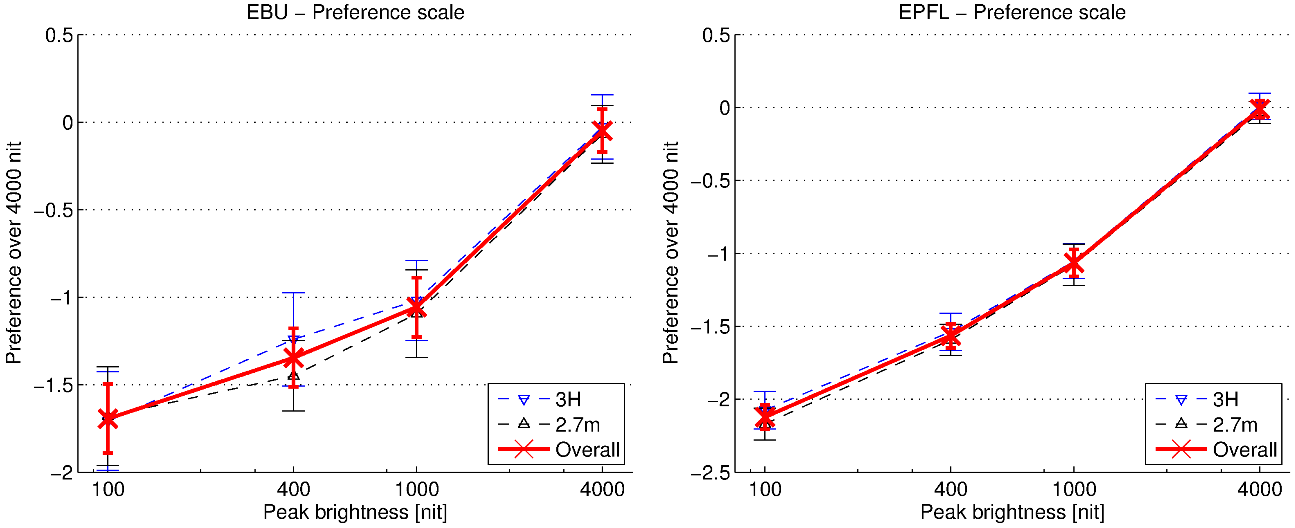


Figure 5: The EBU and EPFL results for preference scale scores

CONCLUSIONS

Scores obtained for the EBU and the EPFL tests show similar trends and lead to similar conclusions. A more detailed statistical analysis of the values for the Pearson linear correlation coefficient and the Spearman rank order correlation coefficient, between the two tests results in values 0.95 and 0.89, respectively.

As a conclusion, horizontal preference scale seems to be appropriate as scoring method. When compared to the forced choice method, which also provides valid results, the preference scale shows a higher accuracy in the confidence intervals and therefore is a better alternative.

The results have also shown that quality differences of brighter HDR content are visually recognized independent from the viewing distance.

Further tests need to be conducted in regard to the ambient light, as this parameter was not considered as a variable in the design of tests.

REFERENCES

1. Seetzen, H., Whitehead, L. A., and Ward, G., 2003. A High Dynamic Range Display Using Low and High Resolution Modulators. SID Symposium Digest of Technical Papers. May, 2003. vol. 34. no. 1. pp. 1450 to 1453.

2. Seetzen, H., Heidrich, W., Stuerzlinger, W., Ward, G., Whitehead, L., Trentacoste, M., Ghosh, A., and Vorozcovs, A., 2004. High dynamic range display systems. ACM Transactions on Graphics. August, 2004. vol. 23. no. 3. pp. 760 to 768.

3. ITU-R BT.500-13, 2012. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union. January, 2012.

ACKNOWLEDGEMENTS

The authors would like to thank Dolby Laboratories Inc. staff, and in particular Suzanne Farrell, Sherif Gallab, and Timo Kunkel for providing Dolby Research HDR RGB backlight dual modulation display (aka Pulsar), HDR video, and for help in the design, implementation, as well as analysis of results.

Furthermore, valuable help and input from the following individuals in the design, implementation and analysis of tests reported in this paper is acknowledged: Andy Quested (BBC), Richard Salmon (BBC), Dagmar Driesnack (IRT) and Giorgio Dimino (RAI).