Subjective Quality Evaluation of High Dynamic Range Video and Display for Future TV

By Philippe Hanhart, Pavel Korshunov, Touradj Ebrahimi, Yvonne Thomas, and Hans Hoffmann

The main objective of this paper is to verify test methodologies for the assessment of high dynamic range (HDR) video. To achieve this, a next generation HDR monitor by Dolby Laboratories was used to display professionally produced HDR content. Two complementary approaches for subjective assessment of HDR video were then designed and carried out at the European Broadcast Union and Ecole Polytechnique Fédérale de Lausanne premises. Results obtained from both evaluations were highly correlated, which shows that they offer a good degree of reliability and reproducibility in different situations. Analysis of the scores in both cases also shows 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 4000 nits, with intermediate improvements at 400 and 1000 nits.

Keywords: *computer displays, video signal processing, high-definition television, optical, image and video signal processing, display technology*

INTRODUCTION

Several technological revolutions have affected the television (TV) industry over the last few decades, such as the shifts from black and 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, three-dimensional television (3DTV) was advertised as the future of television. However, because of low picture quality and the need to wear cumbersome glasses, 3DTV has not yet fulfilled customer expectations. The momentum behind ultrahigh-definition TV (UHDTV) is quickly building, especially during the last couple of years, but some believe that UHDTV could risk experiencing the same reaction as 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 greater 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 effects of these enhancements on viewers' quality of experience? This paper attempts to begin answering this question by proposing new quality evaluation methodologies for HDR video. In fact, efficient and accurate 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 red, green, and blue (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 scenarios, two alternative quality assessment methodologies were designed to obtain highly accurate and reliable measures of perceptual preferences. The first evaluation methodology, performed at the European Broadcast Union (EBU) premises, relied on expert subjects and was carried out using time sequential comparisons, while the second was carried out at the Ecole Polytechnique Fédérale de Lausanne (EPFL) by naïve viewers in simultaneous comparisons (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 this work 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' color vision was checked using standard Ishihara and Snellen vision tests. The subjects who did not pass the vision check (e.g., color blind, 20/20 in Snellen performance) 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

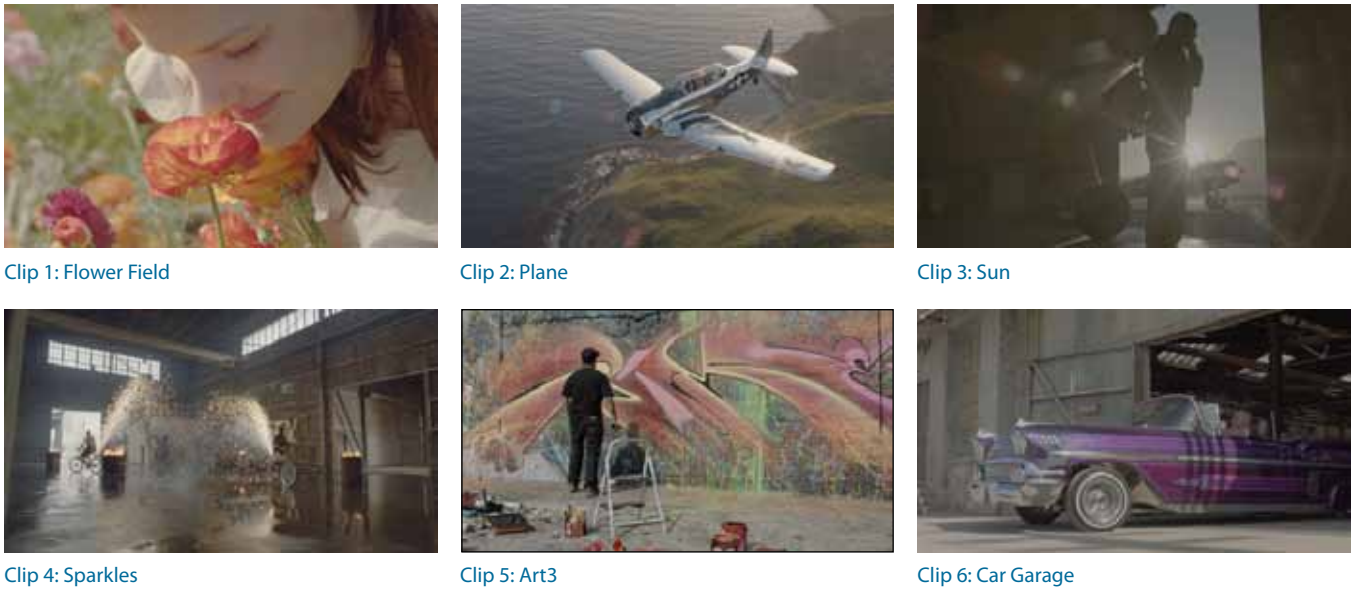

Figure 1. Representative frames of the sequences. Clips 1 to 6 were used in both tests. One additional Hollywood movie clip used in the EPFL test is not included here due to its copyright. The 100 nit versions of the images are shown here because the higher dynamic range images cannot be shown on standard dynamic range displays.

tests. A training session was also organized during which oral instructions were provided to subjects to explain their task and allow them to familiarize themselves with the assessment procedure.

The video material was specifically tailored for display at 100, 400, 1000, and 4000 nits. These four dynamic range levels were selected to be representative of key use cases, as opposed to being uniform perceptual distances. The maximum luminance of 100 nits corresponds to standardized reference monitor levels used for studio color grading and broadcast calibration,[a] 400 nits correspond to max luminance levels in current high-quality liquid crystal display (LCD) and light-emitting diode backlit LCD TVs, 1000 nits correspond to the highest levels found today in consumer TVs, and 4000 nits corresponds to the max luminance level of the HDR display used in this experiment. The displayed video was constructed as follows:

1) 4000 nit version: manually graded by professional colorists from the original video, captured using one of the highest dynamic range digital video cinema cameras (Arri Alexa with 14 *f*-stop range)
2) 1000 nit version: tone-mapped from manually graded 4000 nit version
3) 400 nit version: tone-mapped from manually graded 4000 nit version
4) 100 nit version: tone-mapped from manually graded 4000 nit version.

a. It is 100 nits in the United States and Japan and 120 nits in the EU.

For the 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 perform enhancements or bias importance to specific image regions (as often occurs in human-guided color grading). The input was 12 bits per color in a domain that has characteristics of gamma and log nonlinearities, as suited for HDR.[3] The video sequences were uncompressed. The combination of high bit depth and uncompressed video is intended to remove secondary issues of dynamic range effects on needed bit depth and compression algorithm parameters because the study's aim was to isolate the question of range, starting with maximum luminance.

Evaluations carried out at the EBU relied on expert viewers. A time sequential presentation with both a forced-choice preference, as well as a horizontal preference scale, was conducted, using a 4000 nit graded content as a hidden reference. The hidden reference was shown in every paired comparison, with a randomized position. This methodology was derived from a well-known video comparison methodology.[4] 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,[4] provides a finer comparison of the two conditions. Two viewing distances were tested: 3 H (~1.5 m), a recommended distance for subjective tests, and 2.7 m, a typical distance from TV set in a home environment. The display size was 42 in. diagonal, giving a horizontal field of view (FOV) to the viewers of 33° and 20° for the 1.5 m and 2.7 m distances, respectively. 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

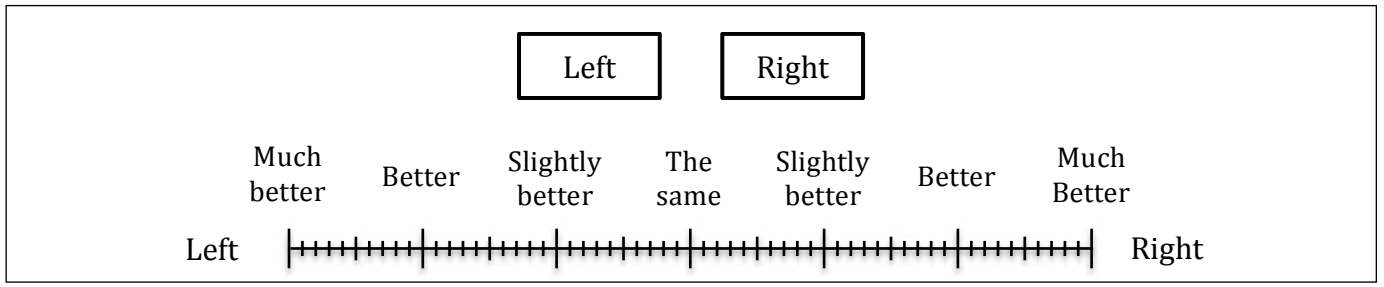

Figure 2. Scoring scales used in the tests (in the EBU tests "Left" was replaced by "A" and "Right" by "B").

simultaneous presentation. These also employed a forced-choice response as well as an horizontal preference scale.

In the EBU tests, six test sequences (**Fig. 1**) were presented in 1080p resolution and a length of 20 sec in a time sequential presentation. In the EPFL tests, in addition to the six sequences mentioned earlier, two more sequences were also included, presented in 1080p resolution and a length of 20 sec in simultaneous side-by-side presentations.

In the EBU tests, the illumination surrounding the display was set to 10 lux for most test groups, and then to 24 lux for one test group to test the effects of ambient illumination levels affecting viewers' light adaptation as well as screen reflectivity consequences. In the EPFL tests, the illumination surrounding the display was set to 20 lux 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 **Fig. 2**. In the EBU tests "Left" was replaced with "A" and "Right" with "B" because of the time sequential display mode used.

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, and presentation of light sources appearing in the scene.

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

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

T1 = 20 sec Test sequence A

T2 = 3 sec Midgray

T3 = 20 sec Test sequence B

T4 = 5 sec Midgray

For the EPFL tests, each video sequence was shown only once. The side-by-side presentation of video sequences resulted in a smaller FOV for each tested video (about half), i.e., 17° and 10° for the 3 H (1.5 m) and 2.7 m viewing distances, respectively.

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 to 4000 nits, to reach 0.5, which corresponds to the preference of 4000 nit sequences (i.e., the hidden reference), which should be theoretically random (50%), since both of the paired comparison video sequences were identical. These results show that there is a significant preference toward 4000 nit displayed content when compared to other alternatives considered in the tests. The relatively flat region from 100 to 1000 nits as compared to the steep rise for 4000 nits indicates the 4000 nit version was substantially different and preferred. It does not necessarily mean that the 100, 400, and 1000 nit versions were close together in appearance and preference. That question requires the second portion of the study, the preference scale responses.

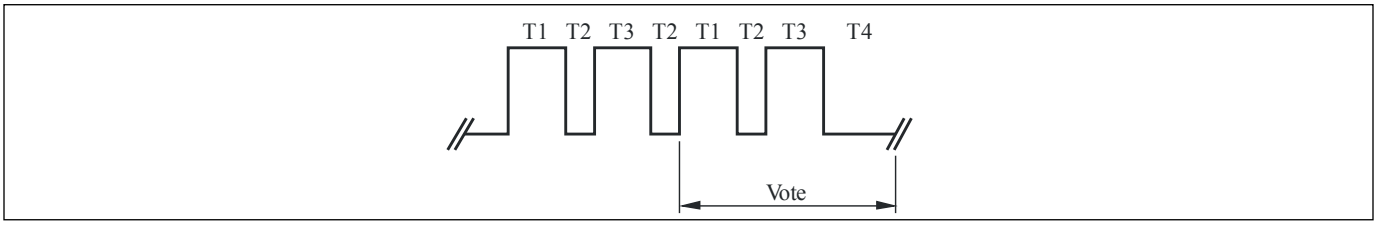

Figure 3. Time sequential mode presentation at the EBU tests.

Figure 5 reports the overall results obtained at the EBU and the EPFL tests for preference scores to the 4000 nit reference and their respective confidence intervals. Results confirm the same trend as in the forced-choice case, but with a better quantification of the steps in terms of preference for the intermediate peak brightness contents. In particular, the tighter confidence intervals in the EPFL tests show that a side-by-side comparison approach not only provides more reliable scores but also better quantify the differences between 400 and 1000 nits in terms of preference when compared to the 4000 nit reference.

CONCLUSION

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

As a conclusion, a horizontal preference scale seems to be appropriate as a 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.

One key question was the use of side-by-side versus sequential comparisons. The sequential comparisons have the drawback of requiring a larger cognitive load on the viewer, due to visual short-term memory (VSTM) limits. On the other hand, while the side-by-side minimizes the demand on VSTM, because both comparisons could be seen with a shorter delay, limitations are still found due to iconic visual memory (<1 sec durations). While the iconic visual memory has higher spatial capacity than the VSTM, the side-by-side comparisons had to crop the video frame so that both could fit in the same full HD display. As a result, the FOV is reduced from 33° to 17° for the closer distance and from 20° to a mere 10° for the farthest viewing distance tested. The FOV is important for HDR display because of the glare due to the optics of the eye and the long tails of its point spread function. Also, other issues are found, such as light adaptation, anchoring, and eye movement aspects that favor a larger FOV presentation. The results showed that these FOV disadvantages of the side-by-side approach were outweighed by the disadvantages of the sequential approach, such as limited VSTM.

The results have also shown that quality differences of brighter HDR content are visually recognized as independent from the viewing distances studied. Nevertheless, this does not mean that viewers had the same experience at 1.5 m than at 2.7 m, as the FOV is known to impact immersiveness, which is one of the factors that contribute to quality of experience. Considering the different FOV aspects, the closer viewing distance is expected to provide a better experience.

Further tests need to be conducted regarding the ambient light because this parameter was not considered as a variable in the design of tests. Other future work involves exploring the shape of the resulting preference scale shown in **Fig. 4** and its difference between the two methods. The shape from the side-by-side study is nearly a straight line in terms of log luminance, which suggests that aspects of Weber's law are in play, possibly related to the hidden reference of 4000 nits (well into the Weber law region) acting as an upper anchor. The shape from the sequential study is more of an expansive nonlinearity, suggesting the 4000 nits are not serving as strongly as an upper anchor. Nevertheless, both tests with the hidden reference showed that higher peak luminance level is preferred.

One way to better understand the impact of the hidden reference is to do a full paired comparison study, where all the maximum lumi-

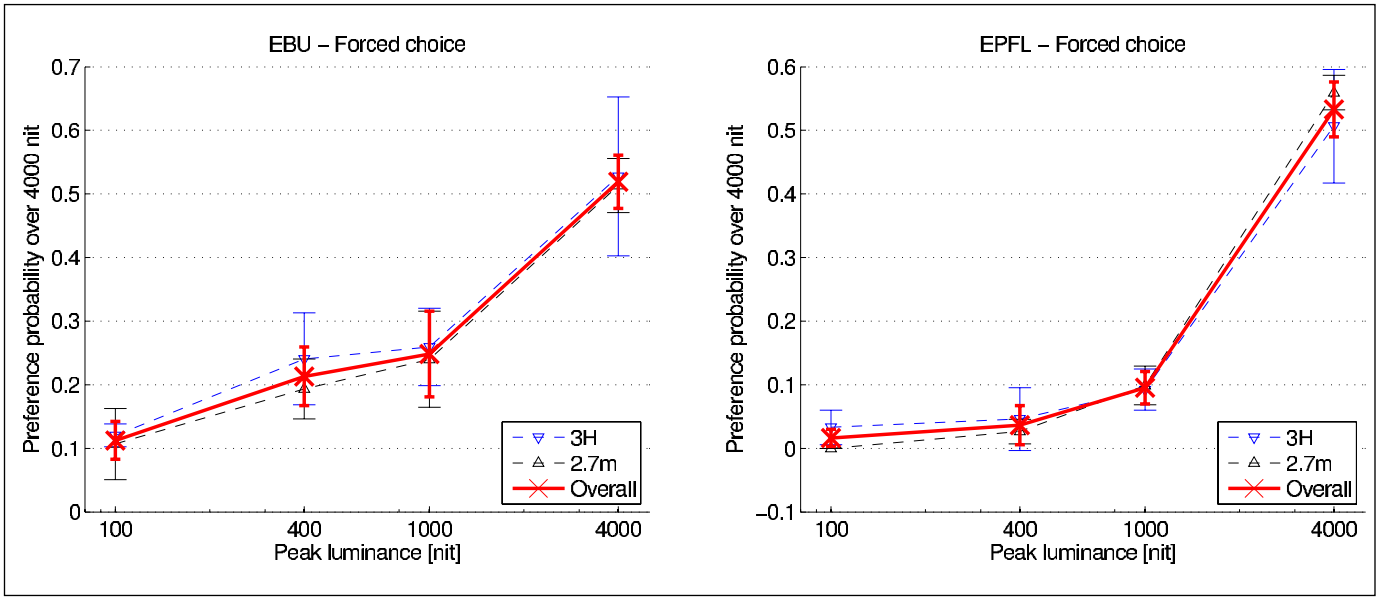

Figure 4. The EBU and EPFL results for forced-choice scores.

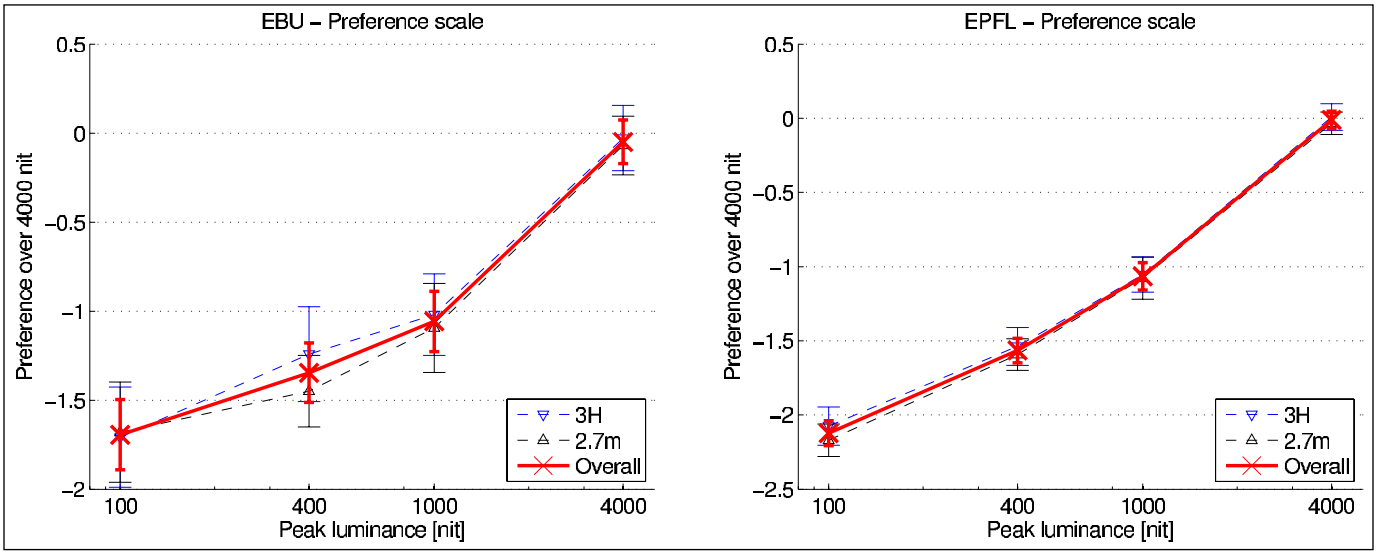

Figure 5. The EBU and EPFL results for preference scale scores.

nance parameters are compared to each other, creating multiple anchors throughout the range, and giving the results in the form of a full interval scale. When considering the standard peak luminance, i.e., 100 nits, as reference, results have shown that 1000 nits is preferred by most viewers.[5] However, to represent highlights, Daly *et al*.[6] found that 18000 nits might be necessary for some images, while 4000 nits is sufficient for most images and that 90% of the viewers are satisfied with a peak luminance of 2400 nits. Therefore, considering these results and the different methodological constraints, a peak luminance between 1000 and 4000 nits leads to a balanced trade-off for most practical purposes.

REFERENCES

1. H. Seetzen, L. A. Whitehead, and G. Ward, "A High Dynamic Range Display Using Low and High Resolution Modulators," *SID Symp. Digest Tech. Papers*, 34(1):1450-1453, May 2003.
2. H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs, "High Dynamic Range Display Systems," *ACM Trans Graphics*, 23(3):760-768, August 2004.
3. S. Miller, M. Nezamabadi, and S. Daly, "Perceptual Signal Coding for More Efficient Usage of Bit Codes," *SMPTE Mot. Imag. J.*, 122(4):52–59, May-June 2013.
4. ITU-R BT.500-13, "Methodology for the Subjective Assessment of the Quality of Television Pictures," International Telecommunication Union, January 2012.
5. P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective evaluation of higher dynamic range video," *Proc. SPIE 9217*, Applications of Digital Image Processing XXXVII, September 2014.
6. S. Daly, T. Kunkel, X. Sun, S. Farrell, and P. Crum, "Preference limits of the visual dynamic range for ultra high quality and aesthetic conveyance," *Proc. SPIE 8651*, Human Vision and Electronic Imaging XVIII, March 2013.

ACKNOWLEDGMENTS

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 help in the design, implementation, and analysis of the results.

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

First presented at IBC 2014, Amsterdam, The Netherlands, 11-15 September 2014. This paper is published here by kind permission of the IBC.

THE AUTHORS

Philippe Hanhart received B.Sc. and M.Sc. degrees in electrical engineering from the Swiss Federal Institute of Technology in Lausanne (EPFL), in 2009 and 2011, respectively. Currently, he is working toward a Ph.D. degree and is a research assistant in the Multimedia Signal Processing Group. He is a recipient of the Anna Barbara Reinhard Prize for Student Excellence from the Institution of Engineering and Technology, top 10% best paper awards in MMSP2014 and ICIP2014, and has authored more than 30 publications. His research interests are in the fields of image and video processing to measure and improve quality of experience in immersive multimedia technologies, such as 3D, ultrahigh-definition, and high dynamic range (HDR).

Pavel Korshunov is a postdoctoral researcher in the Multimedia Signal Processing Group at EPFL. He received a Ph.D. in computer science from National University of Singapore (NUS). He was a recipient of ACM TOMCCAP Nicolas D. Georganas Best Paper Award in 2011, two top 10% best paper awards in MMSP 2014 (16th International Workshop on Multimedia Signal Processing), and the top 10% best paper award in ICIP 2014 (International Conference on Image Processing. Korshunov is a co-editor of the new JPEG XT standard for HDR images, and has authored more than 50 publications. His research interests include computer vision and video analysis, video streaming, video and image quality assessment, crowdsourcing, high dynamic range imaging, ultrahigh-definition imaging, focus of attention, visual privacy protection mechanisms and their evaluation.

Touradj Ebrahimi is a professor at the EPFL, where he is active in teaching and research of multimedia signal processing. He is also the convener of the JPEG committee, which has produced a family of standards that have revolutionized the world of imaging. He represents Switzerland as the head of its delegation to MPEG and JPEG standards and has contributed technical solutions to a large number of JPEG and MPEG standards and particularly in MPEG-2, MPEG-4, AVC/H.264, MVC, 3DVC, and HEVC/H.265. Ebrahimi serves as consultant, evaluator and expert for European Commission projects, for various governmental funding agencies in Europe, and also advises a few venture capital companies in Switzerland as a scientific and technical auditor.

Yvonne Thomas graduated with a degree in television technologies and electronic media engineering from the University of applied Science Wiesbaden (HSRM), Germany, in October 2010. She received a prominent award of the ARD/ZDF Academy for her thesis at the IFA in Berlin in September 2011. Following these studies, she joined the European Broadcasting Union (EBU) in the Technology and Innovation department in 2011. Since then she has been responsible for internal and external projects on 3D and future television technologies, such as UHDTV and LED studio lighting. Thomas coordinates the EBU's Strategic Program BeyondHD in which she led the creation of UHDTV and 3D test content. These were unique experiences and early bird projects that helps EBU members and research institutes to conduct further tests on formats beyond HD (codecs, displays, compression etc.). Further to this, Thomas has been involved in several standardization bodies, such as Digital Video Broadcasting and SMPTE, building the basis for more technical and strategic discussions and decision making for UHD technologies.

Hans Hoffmann is senior manager and head of media fundamentals and production technologies in the EBU Technology and Innovation department. He joined the Institut für Rundfunktechnik (IRT) in 1993 as a research staff in new television production technologies. In February 2000, he became a senior engineer at the EBU, where he has been leading many activities on media, production technologies, video codec evaluations, established the EBU HDTV testing lab, and IT-based digital workflows and recently UHDTV. He has authored many EBU technical documents and IEEE (Institute of Electrical and Electronics Engineers) papers and is a standing speaker and contributor to international conferences. A SMPTE Fellow, Hoffmann served as SMPTE Engineering Vice President from 2011-2013. He is also a member of the FKT and IEEE.