Visual Saliency Aided High Dynamic Range (HDR) Video Quality Metrics

Amin Banitalebi-Dehkordi[1], Maryam Azimi[1], Mahsa T. Pourazad[1,2], and Panos Nasiopoulos[1]
[1]University of British Columbia (UBC), Vancouver, BC, Canada
[2]TELUS Communication Inc., Canada
Corresponding author: dehkordi@ece.ubc.ca

***Abstract*— As High Dynamic Range (HDR) video is emerging as the next revolution in digital entertainment, finding effective ways for measuring its visual quality is of paramount importance. In this paper, we utilize the saliency information derived from an HDR Visual Attention Model (VAM), called LBVS-HDR, for assessing the quality of HDR video content. To this end, this saliency information is incorporated into existing state-of-the-art HDR quality metrics such as the HDR-VDP-2, deltaE2000, mPSNR, and tPSNR as well as the Standard Dynamic Range (SDR) benchmark quality metric, PSNR. The Visual Information Fidelity (VIF) index is also included in our comparisons, as it is reported to perform well for HDR content. Comparing the results of the VAM-aided quality metrics with those of the original ones, we verified that, in general, using saliency prediction for HDR quality assessment improves the performance of all the existing quality metrics. We also observed that the VIF index achieves the highest correlation between the objective and subjective test results.**

Keywords—visual attention model, saliency prediction, High Dynamic Range, quality assessment

I. Introduction

High Dynamic Range (HDR) content has recently received significant recognition in several multimedia application areas as it delivers a dynamic range close to what is perceived by the human visual system (HVS). The human eye can perceive a dynamic range of about 5 orders of magnitude (difference in power of 10 between two values) in real life scenes at a single adaptation time [1]. We should note here that the dynamic range of the existing legacy consumer displays is only about 2 to 3 orders of magnitude. Higher dynamic range video comes at a price of larger amounts of data that need to be captured, transmitted, and displayed.

In practice, HDR images are captured as the amount of light captured by a camera, represented by floating point values. However, depending on the HDR file format used, the floating point HDR values can be converted to 10-16 bits integer values [2-4]. To reduce the size of video data and meet the bit-rate requirements of broadcasting or disc storage-capacity limitations, video compression techniques are utilized. It is, therefore, very important to have efficient HDR compression schemes that will facilitate storage and delivery of such large amounts of data [5]. The user-end viewing experience depends on how well the visual quality is preserved during the HDR video delivery pipeline. As subjective quality evaluation is not a practical option in in real-time applications, objective quality metrics are employed to assess the quality of the HDR content at acquisition, transmission, and display [6-7]. The first attempts for HDR quality assessment included the use of SDR quality metrics to evaluate the quality of tone-mapped versions of the HDR image at different exposure levels [8]. The overall quality in this case (known as multi-exposure method) is estimated as the average of quality metric values from different exposures. Another approach by Aydin et al. [9] proposes to bring the dynamic range of the HDR content to a lower non-redundant range supported by the SDR metrics. This method is known as Perceptually Uniform (PU) encoding approach. There are also several PSNR-based HDR metrics proposed in the literature that basically modify the formulation of Mean Square Error (MSE) to account for the higher dynamic range of the signals. Among these metrics, tPSNR attempts to remove the bias towards any particular color transfer function [10-12]. The deltaE2000 is also another PSNR-based metric that assesses the color difference according to CIE DE 2000 standard [10-12]. In addition to the above-mentioned metrics, there exists another class of HDR metrics, which are independent of the dynamic range of the content. HDR-VDP-2 is an example of such class of metrics [13].

The performance of various HDR and SDR quality metrics for quality evaluation of compressed HDR video content has been investigated in [6,7,10]. However, none of these studies incorporates an HDR Visual Attention Model (VAM) to identify visually important regions of HDR content nor they use an efficient pooling mechanism for quality assessment. In this paper, we investigate the added value of utilizing the saliency information on existing metrics when measuring the quality of compressed HDR video content. In our approach, the saliency of HDR videos is predicted by a Learning-Based Visual Saliency (LBVS) prediction model [14]. This LBVS-HDR saliency predictor is designed based on HDR saliency feature extraction and learning-based feature fusion. The saliency maps predicted by LBVS-HDR are then utilized in the formulation of the quality metrics used in the HDR video compression standardization activities [10]. The quality metrics we used include mPSNR (multi-exposure PSNR), tPSNR,

This work was partly supported by Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant STPGP 447339-13 and the Institute for Computing Information and Cognitive Systems (ICICS) at UBC.

deltaE2000, and HDR-VDP-2. In addition, we utilized the saliency maps predicted by LBVS-HDR in the formulation of PSNR (to use as a benchmark) and in the Visual Information Fidelity (VIF) index [15], as VIF has been reported to perform well for HDR quality assessment [10]. To investigate the efficiency of the proposed scheme, we measure the correlation between the results of the saliency-aided quality metrics with the subjective test results and compare it with the correlation between the quality metrics and the subjective test scores over a representative HDR video dataset.

The rest of this paper is organized as follows: Section II describes how saliency information is integrated in each HDR metric, Section III explains the procedure of preparing the HDR video dataset and the test setup, Section IV contains the results and discussions, and Section V concludes the paper.

II. Integrating Saliency Information in HDR Metrics

This section provides a brief description on how the saliency information predicted by the LBVS-HDR model is integrated into different quality metrics. Using the LBVS-HDR model, one saliency map is generated per each HDR video frame. The saliency map is essentially a gray-scale image of the same size as the input frame. Each pixel of this saliency map has a value between 0 and 1 that represents the likelihood of that pixel to be watched by a viewer [14]. The saliency maps are extracted from the original uncompressed HDR content, as this results in more accurate saliency prediction. In our method, we use the saliency map values as weights to local metric measurements, which could be a measure of visible distortions or similarities in the spatial or frequency domain and possibly at different scales (see Fig. 1). The saliency maps have been made publicly available to the research community to facilitate further research and development on HDR video [16]. In the following subsections we elaborate on how the predicted visual saliency maps are integrated into each HDR metric.

A. *LBVS-HDR Aided PSNR (PSNR$_S$)*

PSNR is the most common objective quality metric used for SDR video compression applications. In our study, we employ the saliency maps predicted by LBVS-HDR as a weighting function for the measured error, i.e., mean square error (MSE), as follows:

$$PSNR_S = E_t\left\{10\log\left(\frac{255^2}{MSE_S(t)}\right)\right\} \quad (1)$$

$$MSE_S(t) = E_{x,y}\left\{\left|I(x,y,t) - I'(x,y,t)\right|^2 \times S(x,y,t)\right\}$$

where t is the timestamp (frame number), E_t is the temporal averaging operand, MSE$_s$ is the modified version of MSE, x and y denote the pixel coordinates, I and $I^{'}$ denote the reference and compressed frames, S is the saliency map, and $E_{x,y}$ denotes the spatial averaging operand.

B. *LBVS-HDR Aided VIF (VIF$_S$)*

VIF is a relative measure of mutual information between the perceived compressed and reference signals and their original counterparts [15]. The pixel-wise LBVS-HDR aided implementation of VIF is formulated as follows:

$$VIF_S = E_t\left\{\frac{\sum_{m=1}^{M}\left\{\frac{1}{2}\sum_{i,k}\log\left(1+\frac{g^2 s_{i,m}^2(t)\lambda_{k,m}(t)}{\sigma_v^2+\sigma_n^2}\right)S_m(i,k,t)\right\}}{\sum_{m=1}^{M}\left\{\frac{1}{2}\sum_{i,k}\log\left(1+\frac{s_{i,m}^2(t)\lambda_{k,m}(t)}{\sigma_n^2}\right)S_m(i,k,t)\right\}}\right\} \quad (2)$$

where i and k are spatial subband indices, s_i, λ_k, σ_n, and σ_v are VIF parameters (see [15] for more details), M is the total number of subbands that the frames are decomposed into, and S_m is the saliency map resized to the scale of subband m.

C. *LBVS-HDR Aided HDR-VDP-2 (HDR-VDP2$_S$)*

HDR-VDP-2 is an HDR image quality metric, which is designed to assess the quality in all luminance conditions. This metric utilizes steerable pyramids [17] to perform a multiscale decomposition of an HDR image. For each band, a perceptually linearized contrast difference is calculated. Multiband pooling on the difference values followed by applying a logistic function, results in an overall quality index. We use the saliency information in each band, and at different scales to weight different regions of the HDR image according to their visual importance. The LBVS-HDR aided HDR-VDP-

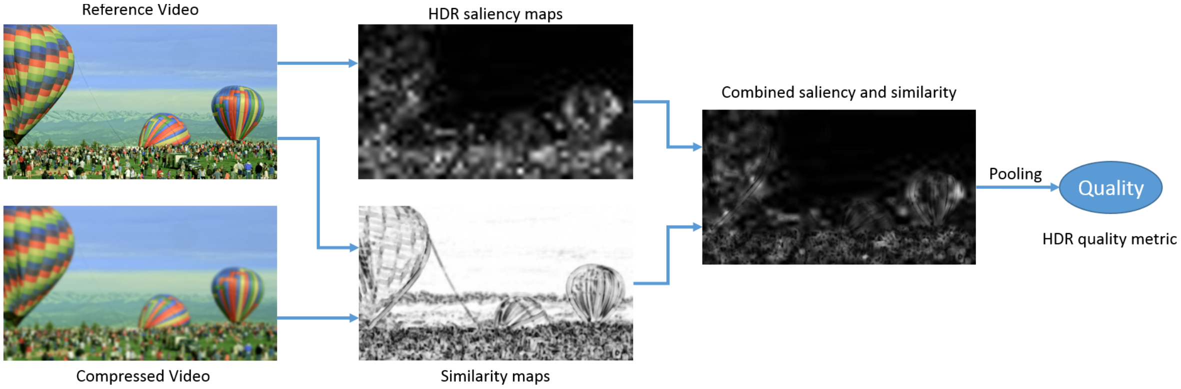

Fig. 1. Saliency inspired quality assessment of HDR video

2 for an HDR video is formulated as:

$$VDP2_S = E_t\left\{\frac{1}{1+e^{q_1(Q_S(t)+q_2)}}\right\} \tag{3}$$

where E_t is the temporal averaging operand, q_1 and q_2 are constant logistic parameters, and Q_S is defined as:

$$Q_S(t) = E_{f,o}\left\{w_f \log\left(E_{x,y}\left\{D^2[x,y,f,o,t] \times S(x,y,f,o,t)\right\}\right)\right\} \tag{4}$$

where *x,y* denote the pixel coordinates, *D[f,o]* denotes the contrast difference for the *f*th spatial frequency band and the *o*th orientation of the steerable pyramid, *S(f,o)* denotes the saliency map rescaled (by resizing) corresponding to the band *(f,o)*, w_f is a weighting constant, and *E* is the averaging operand.

D. LBVS-HDR Aided mPSNR (mPSNR$_S$)

Multi-exposure *PSNR* (or *mPSNR*) is designed to evaluate the *PSNR* value at various exposure levels. This metric is performed over the linear RGB values (16bit EXR files) and incorporates gamma curves for individual color channels. The saliency aided $mPSNR_s$ is evaluated as follows:

$$mPSNR_S = E_t\left\{10\log\left(\frac{255^2}{mMSE_S(t)}\right)\right\} \tag{5}$$

where E_t is temporal averaging operand and $mMSE_S$ is defined as:

$$\begin{aligned} mMSE_S(t) = E_c\Big\{E_{x,y}\Big\{(&\left|R(x,y,c,t)-R'(x,y,c,t)\right|^2 \\ &+\left|G(x,y,c,t)-G'(x,y,c,t)\right|^2 \\ &+\left|B(x,y,c,t)-B'(x,y,c,t)\right|^2)\times S(x,y,t)\Big\}\Big\} \end{aligned} \tag{6}$$

where *c* denotes the exposure level, and *R*, *G*, and *B* represent the three color channels.

E. LBVS-HDR Aided tPSNR (tPSNR$_s$)

The tPSNR metric utilizes an averaging method to avoid biasing towards a particular color transfer function. This metric computes the MSE value between the three color channels, which are derived from averaging the PQ_TF and Philips_TF curves [11]. To this end, the content is first converted to the linear HDR format, and then it is converted to *XYZ* space. Each of the *X*, *Y*, and *Z* components are transferred using the mentioned two transfer functions and the Sum of Squared Error (SSE) is computed for each color component. The average SSE results in an overall error value, from which a PSNR is calculated. We incorporate the HDR saliency maps as weights (element-wise) to the color components, after the transfer functions are applied to them. As a result, the SSE values are calculated based on the saliency-weighted color channels (similar to (6) but for *X*, *Y*, and *Z*).

F. LBVS-HDR Aided deltaE2000

The *deltaE2000* metric is a quantitative measure of color difference introduced by the CIE in 2001. A PSNR-like variation of this metric was considered for the MPEG HDR quality assessment activities [10-11]. Here, the content is first converted to 4:4:4 linear EXR format and a *deltaE* value is computed for each pixel. We incorporate the HDR saliency maps to put emphasis on the *deltaE* values associated with the pixels of higher visual importance to the human eye as follows:

$$deltaE2000_{PSNR} = E_t\left\{10\log\left(\frac{10000}{E_{x,y}\{DE(x,y,t)\times S(x,y,t)\}}\right)\right\} \tag{7}$$

where $E_{x,y}$ evaluates the average color difference over a distortion specific window and *DE* is actual CIE *deltaE2000* color difference value [11].

III. Experiment Setup

This section provides details on the experiment set up and the subjective tests performed to investigate the performance of the saliency-aided quality metrics.

A. Video Data Preparation

We used four video sequences of the HDR videos provided by Technicolor and CableLabs to the MPEG community for our experiment [18] (see Table 1). In order to encode the original half floating point HDR video content, we follow the workflow shown in Fig. 2. The 10-bit HDR videos are encoded at four different QP levels using the latest HEVC encoder software, HM 16.2. The QPs used for the Tibul2 video are the ones recommended in [6], and for the videos Market3, FireEater2, and BalloonFestival the values are {29, 33, 37, 41}, {21, 25, 29, 33}, and {18, 26, 34, 38}, respectively. The reason for the introducing new QPs is that the lowest QP value recommended by MPEG for these two videos did not result in noticeably different visual quality levels when viewed on a SIM2 display. The random access high efficiency (RA-HE) configuration of HEVC was used to ensure achieving the highest compression performance [5]. Input, internal, and output bitdepths were all set to 10.

B. Display

A full HD 47” SIM2 HDR LCD display (peak luminance of 6000 cd/m^2) with individually controlled LED backlight

Table 1 HDR video database

Sequence	Resolution	Frame Rate (fps)	Number of Frames	Scene Type	Cropped Area
FireEater2	1920×1080	25	200	Outdoor/Night	550 - 1497
Market3	1920×1080	50	400	Outdoor/Day light	100 - 1047
Tibul2	1920×1080	30	240	Computer-generated	800 - 1747
BalloonFestival	1920×1080	24	250	Outdoor/Day light	1-948

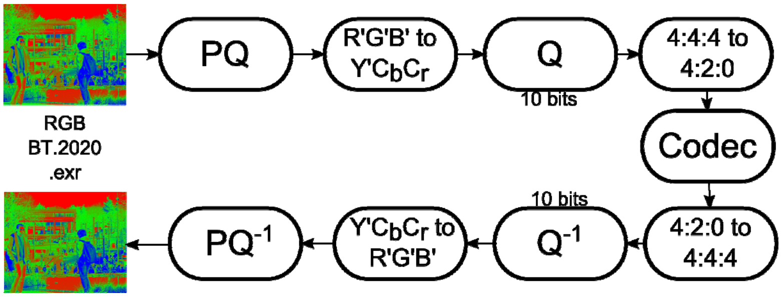

Fig. 2. The process of HDR video compression

modulation was used in our subjective tests. Prior to the experiments, the monitor was calibrated to ensure linear transfer responses for each color channel. The decoded video sequences were converted into OpenEXR frames and then into display-specific bitmap format. The SIM2 display supports BT.709 [19] gamut, while the HDR video sequences provided by MPEG [18] are in BT.2020 container [20], although the gamut of those sequences is not exceeding the BT.709 gamut. To incorporate the characteristics of the display gamut, conversion is performed using the HDRTools software [11].

The SIM2 display at HDR mode expects the input color values in 24-bit LogLuv format [4]. For this reason, after gamut conversion, we convert the RGB float values into LogLuv format.

C. *Subjective Tests*

The subjective tests were performed according to the Recommendation BT.500-13, DSIS method [21]. Both the original and stimuli, which is the decoded HDR, are shown to the viewers at the same time in a side-by-side manner. In order to show the videos side-by-side, we had to crop them along their width while keeping the height (1080p) the same, to avoid changing the resolution. Fig. 3 shows the cropped rectangle for each of the contents. The x coordinates of the cropped windows are shown in Table 1. We tried to select a rectangle that contains the most important information and the moving objects.

During the tests, the subjects were aware which one of the videos was the original one and the position of the original video on the screen remained un-changed throughout each test session. However, the test and original videos were swapped over different test sessions to have an unbiased observation. The order of videos in each session of the test was randomized and extra care was taken for the same sequence not to be shown consecutively.

Subjects were asked to compare the quality of the test video with that of the original one and assign a discrete rating scale ranging from 1 being the worst quality to 10 being the best quality matching the original video.

D. *Viewers*

Eighteen adult subjects (10 males and 8 females) with the age range of 24 to 30 participated in our subjective test. Prior to the tests, all subjects were screened for color blindness and visual acuity by the Ishihara chart and the Snellen charts, respectively. Subjects that failed the pre-screening test did not participate in the test. None of the participants were aware of the test objectives. An oral and a written instruction of the test were presented to the subjects prior to the test. In order to familiarize the participants with the test procedure, at the beginning we had a training session including two videos (different from the ones in the actual test) with compression artifacts. All the tests were conducted with three subjects per session.

After collecting the subjective test results, the outlier subjects were detected according to the ITU-R BT.500-13 recommendation. Two outliers, out of 18 subjects, were detected and their input data were discarded from the results.

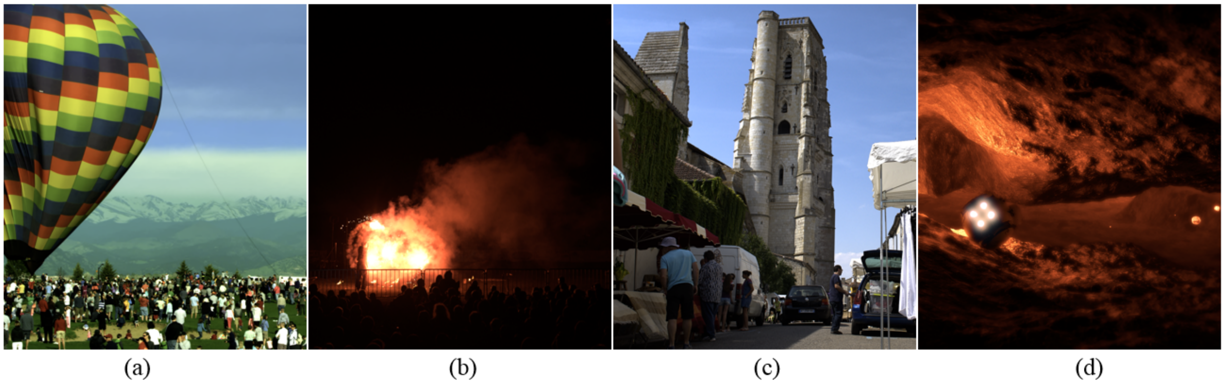
(a) (b) (c) (d)

Fig. 3. Snapshots of the HDR videos, from the left to right: Balloon, Fire-eater, Market, Tibul

Table 2. Statistical performance of different quality metrics with and without integration of the saliency maps

Quality Metric	Performance Metric	PCC		SCC		RMSE	
		Original	With Saliency	Original	With Saliency	Original	With Saliency
PSNR		0.4799	0.5578	0.6269	0.6799	0.4765	0.1514
VIF		**0.6440**	**0.7485**	**0.8035**	**0.8521**	**0.1294**	**0.0947**
HDR-VDP-2		0.4111	0.4359	0.5607	0.6740	0.1749	0.1594
mPSNR		0.3891	0.4051	0.1354	0.1648	0.1772	0.1770
tPSNR		0.2284	0.2889	0.4018	0.4341	0.6288	0.1809
deltaE2000		0.4842	0.5311	0.6269	0.6549	0.4765	0.1634

IV. Results and Discussions

After calculating the mean opinion scores (MOS) and computing the objective quality of the test HDR content using both the original and saliency-aided metrics, a comparison analysis was performed to investigate the performance of the saliency-aided quality metrics. To this end, three different performance metrics are used: 1) the Pearson Correlation Coefficient (PCC) that measures accuracy, 2) the Spearman Correlation Coefficient (SCC) that measures the monotonicity, and 3) the Root Mean Square Error (RMSE) that measures the accuracy of a mapping between an objective metric and MOS. Table 2 shows the performance of different HDR quality metrics in predicting the subjective quality of the compressed videos. Fig. 4 illustrates the improvement achieved through integrating the saliency information to each quality metric (in terms of PCC, SCC, and RMSE). As it is observed, the saliency information has improved the performance of all of the quality metrics used in this experiment. The amount of improvement, however, varies for different metrics. The same results also show that VIF outperforms the other state-of-the-art quality metrics.

V. Conclusions

The objective of this paper was to investigate if using HDR saliency prediction along existing quality metrics could improve the accuracy of measuring the visual quality of HDR content. To this end, we chose a representative set of HDR videos, and encoded them based on the MPEG recommendations for HDR video compression. Then, we used our HDR VAM to extract the saliency information for the HDR videos. Performance evaluations showed that integrating the saliency information in existing state-of-the-art HDR quality metrics increased the accuracy in predicting the visual quality of the compressed HDR videos.

References

[1] J. A. Ferwerda, "Elements of Early Vision for Computer Graphics," *Computer Graphics and Applications*, vol. 21, no. 5, pp. 22–33, 2001.

[2] R. Boitard, M.T. Pourazad, P. Nasiopoulos, and J. Slevinsky, "Demystifying HDR Technology," IEEE Consumer Electronics Magazine, vol. 4, issue 4, pp. 72 – 86, Oct. 2015.

[3] R. Bogart, F. Kainz, and D. Hess, "Openexr image file format," ACM SIGGRAPH, Sketches & Applications 2003.

[4] G. W. Larson, "Logluv encoding for full-gamut, high dynamic range images," Journal of Graphics Tools, vol. 3, no. 1, pp. 15–31, 1998.

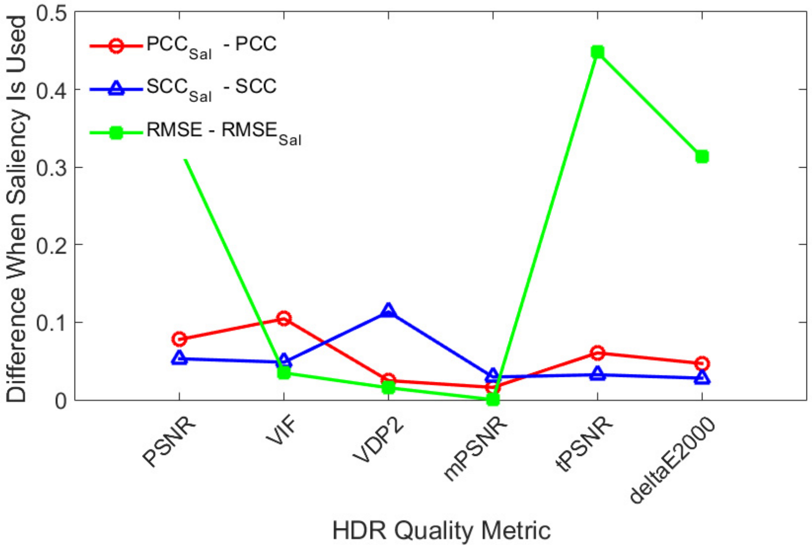

Fig. 4. Resulting improvements when saliency maps of LBVS-HDR [14] VAM are integrated to the HDR quality metrics.

[5] A. Banitalebi-Dehkordi, M. Azimi, M. T. Pourazad, and P. Nasiopoulos, “Compression of high dynamic range video using the HEVC and H. 264/AVC standards,” QSHINE 2014 Conference, Greece, Aug. 2014 (invited paper).

[6] A. Banitalebi-Dehkordi, M. Azimi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, “Quality assessment of High Dynamic Range (HDR) video content using existing full-reference metrics,” ISO/IEC JTC1/SC29/WG11, France, Oct. 2014.

[7] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, “Evaluating the performance of existing full-reference quality metrics on High Dynamic Range (HDR) Video content,” ICMSP 2014: XII International Conference on Multimedia Signal Processing, Nov. 2014, Venice, Italy.

[8] J. Munkberg, P. Clarberg, J. Hasselgren, and T. Akenine-Möller, “High dynamic range texture compression for graphics hardware,” *ACM Transactions on Graphics*, vol. 25, no. 3, p. 698, Jul. 2006.

[9] T. O. Aydin, R. Mantiuk, and H. P. Seidel, “Extending quality metrics to full luminance range images,” *SPIE, Human Vision and Electronic Imaging XIII*, San Jose, USA, Jan. 2008.

[10] M. Rerabek, P. Korshunov, Ph. Hanhart, and T. Ebrahimi, “Correlation of subjective scores and objective metrics for HDR video quality assessment,” ISO/IEC JTC1/SC29/WG11 MPEG2014/ m35273, October 2014, Strasbourg, France.

[11] ISO/IEC JTC1/SC29/WG11, “HDRTools: Software updates,” Doc. M35471, Geneva, Switzerland, February 2015.

[12] Ph. Hanhart, M. Rerabek, and T. Ebrahimi, “Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies,” Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland, July 2015.

[13] Rafal Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. 2011. “HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions”. *ACM Trans. Graph*. 30, 4, Article 40, 14 pages, July 2011.

[14] A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and Panos Nasiopoulos, “A Learning Based Visual Saliency Fusion Model For High Dynamic Range Video (LBVS-HDR),” 23rd European Signal Processing Conference, EUSIPCO 2015.

[15] H. R. Sheikh and A. C. Bovic, “Image information and visual quality”, *IEEE Transactions on Image Processing*, Vol. 15, NO. 2, Feb. 2006.

[16] LBVS-HDR available at: http://dml.ece.ubc.ca/data/LBVS-HDR/

[17] E P Simoncelli and W T Freeman. The Steerable Pyramid: A Flexible Architecture for Multi-Scale Derivative Computation. IEEE Second Int'l Conf on Image Processing. Washington DC, October 1995.

[18] D. Touz´e and E. Francois, "Description of new version of HDR class A and A' sequences," in ISO/IEC JTC1/SC29/WG11 MPEG2014/M35477. Geneva, Switzerland: Feb. 2015.

[19] ITU, "Recommendation ITU-R BT.709-3: Parameter values for the HDTV standards for production and international programme exchange," International Telecommunications Union, 1998.

[20] ITU, "Recommendation ITU-R BT.2020: Parameter values for ultrahigh definition television systems for production and international programme exchange," International Telecommunications Union, 2012.

[21] International Telecommunication Union, "Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service," in Recommendation ITU-R BT.500-13, vol. 13, 2012.