INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2014/m37317
Oct. 2015, Geneva, Switzerland

Source [1]University of British Columbia (UBC) & [2]TELUS Communication Inc.
Status Input
Title Saliency-Aided HDR Quality Metrics
Author Amin Banitalebi-Dehkordi[1], Maryam Azimi[1], Mahsa T. Pourazad[1,2], and Panos Nasiopoulos[1]

Abstract

As we enter the final stages of the standardization efforts for High Dynamic Range (HDR) video compression, the availability of an effective HDR quality metric is becoming more critical. So far, metrics used for measuring the visual quality of HDR content include existing Standard Dynamic Range (SDR) metrics, some metrics designed exclusive HDR and others that are independent of dynamic range.. In this contribution, 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 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.

1 Introduction

The performance of various HDR and SDR quality metrics for quality evaluation of compressed HDR video content has been investigated in [1-3]. However, these previous studies on the HDR quality evaluation metrics fail to incorporate an HDR Visual Attention Model (VAM) to identify visually important regions of HDR content and use an efficient pooling mechanism for quality assessment. To predict the saliency of HDR videos, a Learning-Based Visual Saliency (LBVS) called LBVS-HDR has been proposed utilizing an eye-tracking experiment [4]. The LBVS-HDR is designed based on HDR saliency feature extraction and learning-based feature fusion. In this contribution, we investigate the added value of utilizing the saliency information when the existing HDR quality metrics are used. We extract the saliency maps using LBVS-HDR and utilize the information in the formulation of mPSNR (multi-exposure PSNR), tPSNR, deltaE2000, and HDR-VDP-2, as these metrics are used in the HDR video compression standardization activities [3]. We also use PSNR as a benchmark and the Visual Information Fidelity (VIF) index [5] as the latter was reported to perform well for HDR quality assessment (even though VIF is originally designed for SDR content).

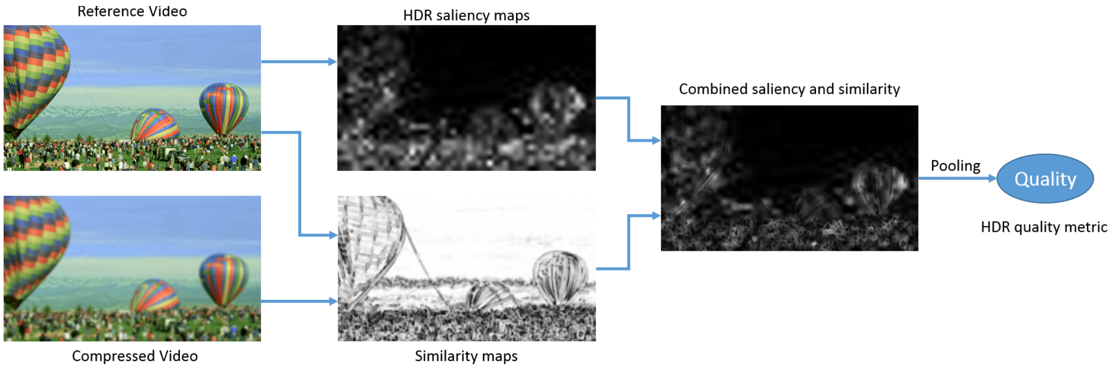

Fig. 1. Saliency inspired quality assessment of HDR video

2 Integrating saliency information in HDR quality 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 frame. The saliency map is essentially a gray-scale image of the same size as the input image. 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 [4]. 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. Figure. 1 shows the flowchart of the proposed method. In the following sub-sections we elaborate on saliency integration for each specific HDR metric.

2.1 *LBVS-HDR aided PSNR (PSNR$_S$):*

PSNR$_S$ is formulated by:

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

where t is the timestamp (frame number), E_t is temporal averaging operand, and MSE$_S$ is the modified version of MSE using the saliency maps as follows:

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

where x and y denote the pixel coordinates, I and I' denote the reference and compressed frames, S is the saliency map, and $E_{x,y}$ denote spatial averaging operand.

2.2 *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 [5]. Pixel-wise LBVS-HDR aided implementation of VIF is formulated as follows:

$$VIF_S = E_t \left\{ \frac{\sum_{m=1}^{M} \left\{ \frac{1}{2} \sum_i \sum_k \log\left(1 + \frac{g^2 s_{i,m}^2(t) \lambda_{k,m}(t)}{\sigma_v^2 + \sigma_n^2} \right) \times S_m(i,k,t) \right\}}{\sum_{m=1}^{M} \left\{ \frac{1}{2} \sum_i \sum_k \log\left(1 + \frac{s_{i,m}^2(t) \lambda_{k,m}(t)}{\sigma_n^2} \right) \times S_m(i,k,t) \right\}} \right\} \qquad (3)$$

where i and k are spatial subband indices, s_i, λ_k, σ_n, and σ_v are other VIF parameters [5], M is the total number subbands that the frames are decomposed into, S_m is saliency map resized to the scale of subband m.

2.3 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 the steerable pyramids [6] to perform a multiscale decomposition on 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 the 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 VDP-2 for an HDR video is formulated as:

$$VDP2_S = E_t \left\{ \frac{1}{1 + e^{q_1 (Q_S(t) + q_2)}} \right\} \qquad (4)$$

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

$$Q_S(t) = E_{f,o}\left\{ w_f \log\left(E_{x,y}\left\{ D^2[x,y,f,o,t] \times S(x,y,f,o,t) \right\} \right) \right\} \quad (5)$$

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.

2.4 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. Saliency aided *mPSNR$_s$* is evaluated as follows:

$$mPSNR_S = E_t \left\{ 10 \log\left(\frac{255^2}{mMSE_S(t)} \right) \right\} \qquad (6)$$

where *mMSE$_S$* is defined as:

$$mMSE_S(t) = E_c\Big\{E_{x,y}\Big\{\big(\left|R(x,y,c,t)-R'(x,y,c,t)\right|^2 + \left|G(x,y,c,t)-G'(x,y,c,t)\right|^2+\left|B(x,y,c,t)-B'(x,y,c,t)\right|^2\big)\times S(x,y,t)\Big\}\Big\} \tag{7}$$

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

2.5 *LBVS-HDR aided tPSNR (tPSNR$_S$):*

tPSNR utilizes an averaging method to avoid biasing towards a particular color transfer function. This metric computes the MSE value between three color channels, which are derived from averaging the PQ_TF and Philips_TF curves [7]. To this end, the content is converted to linear HDR format first (if not already in that format), and then it is converted to *XYZ* space. Each of *X*, *Y*, and *Z* components are transferred using the mentioned two color transfer functions and a Sum of Squared Error (SSE) is computed for each color component. The average SSE results in the 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 will be calculated based on the saliency-weighted color channels (similar to (7) but for *X*, *Y*, and *Z*).

2.6 *LBVS-HDR aided deltaE2000:*

deltaE2000 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 [3,7]. To this end, 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:

$$deltaE2000_{PSNR} = E_t\left\{10\log\left(\frac{10000}{E_{x,y}\{DE(x,y,t)\times S(x,y,t)\}}\right)\right\} \tag{8}$$

where $E_{x,y}$ evaluates the average color difference over a distortion specific window and *DE* is actual CIE *deltaE2000* color difference value [7].

3 Experiment Set up

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

3.1 Video Data Preparation

We used four video sequences of the HDR videos provided by Technicolor and CableLabs to MPEG community for our experiment [8] as reported in Table 1. In order to encode the original half floating point HDR video content, we follow the workflow shown in Figure. 2. The 10-bit HDR videos are encoded at four different QP levels using the latest HEVC encoder software HM

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

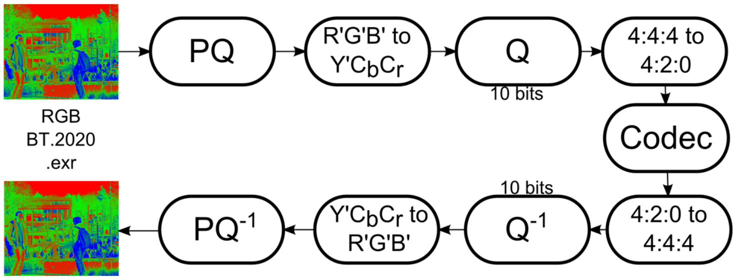

Figure. 2. The process of HDR video compression

16.2. The QPs used for the Tibul2 video are the ones recommended in [1], and for the videos Market3, FireEater2, and BalloonFestival are {29, 33, 37, 41}, {21, 25, 29, 33}, and {18, 26, 34, 38} respectively. ~~The QPs used for each of the videos are the ones recommended in [9] except for the videos BalloonFestival and Market3 for which are coded using QP values of {18, 26, 34, 38}, and {29, 33, 37, 41}, respectively. The reason was that the highest QP recommended by MPEG for these two videos did not achieve a bit-rate low enough to match the lowest bitrate for the other videos.~~ The reason for the introducing new QPs is that the lowest QP recommended by MPEG for these two videos did not result in noticeable 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 [10]. Input, internal, and output bitdepth were all set to 10.

3.2 Display

A full HD 47” SIM2 HDR LCD display (peak luminance of 6000 cd/m^2) with individually controlled LED backlight modulation was used in the 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 [11] gamut while the HDR video sequences provided by MPEG [8] are in BT.2020 container [12], 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 [7].

The SIM2 display at HDR mode expects the input color values in 24-bit LogLuv format [13]. For this reason, after gamut conversion, we convert the RGB float values into LogLuv format.

3.3 Subjective Tests

The subjective tests were performed according to the Recommendation BT.500-13, DSIS method [14]. Both original and the stimuli, which is the decoded HDR, are shown at the same time in a side-by-side manner to the viewers. In order to show the videos side-by-side, we had to crop the videos along their width while keeping the height (1080p) the same, to avoid changing the resolution. Figure. 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

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

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

matching the original video.

3.4 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 was 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.

4 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 is performed to investigate the performance of the saliency-aided quality metrics. To this end, three different performance metrics are used: 1) Pearson Correlation Coefficient (PCC) that measures

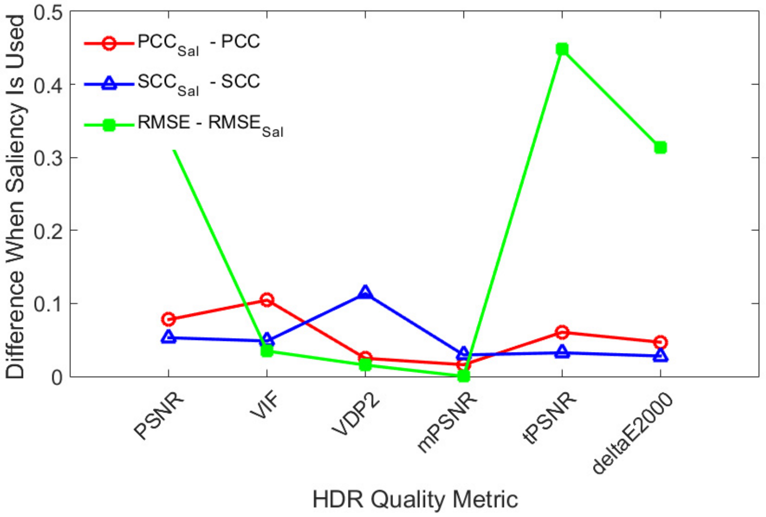

Fig. 4. Resulting improvements when saliency maps of LBVS-HDR [4] VAM are integrated to the HDR quality metrics.

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

the accuracy, 2) Spearman Correlation Coefficient (SCC) that measures the monotonicity, and 3) Root Mean Square Error (RMSE) that measures the accuracy of a mapping between an objective metric, and the MOS. Table 2 shows the performance of different HDR quality metrics in predicting the subjective quality of the compressed videos. Figure. 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.

5 CONCLUSIONS AND RECOMMENDATIONS

Our objective was to investigate if using HDR saliency prediction could improve the performance of existing HDR quality metrics. To this end, we chose a 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. Integrating the saliency information in existing state-of-the-art HDR quality metrics increased their performance in predicting the visual quality of the compressed videos.

We suggest considering HDR saliency prediction for assessing the quality of HDR video.

6 References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] A. Luthra, E. Francois, and W. Husak, "Call for Evidence (CfE) for HDR and WCG Video Coding," in ISO/IEC JTC1/SC29/WG11 MPEG2014/N15083. Geneva, Switzerland: Feb. 2015.

[10] 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).

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

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

[13] L. G. Ward, "LogLuv Encoding for Full-Gamut, High-Dynamic Range Images," Journal of Graphics Tools, vol. 3, no. 1, pp. 15-31, Jan. 1998.

[14] 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.