A Just Noticeable Difference Subjective Test for High Dynamic Range Images

Ayyoub Ahar*‡, Saeed Mahmoudpour*‡, Glenn Van Wallendael†‡, Tom Paridaens †‡, Peter Lambert †‡ and Peter Schelkens*‡
*Vrije Universiteit Brussel (VUB), Dept. of Electronics and Informatics (ETRO), Pleinlaan 2, B-1050 Brussels, Belgium
†Gent University, Gent, Belgium
‡imec, Kapeldreef 75, B-3001 Leuven, Belgium

***Abstract*—High Dynamic Range (HDR) imaging captures a wide range of luminance existing in real-world scenes. Due to large luminance levels and higher brightness of HDR displays, artefacts can be more noticeable to the Human Visual System (HVS). In a first attempt to experimentally quantify those noticeable levels for HDR images, we pioneered in conducting an exhaustive and comprehensive Just Noticeable Difference (JND) subjective experiment of which the outcome is presented in this paper. Six distortions including JPEG, JPEG2000, noise, blur, contrast change, and quantization artefacts have been considered in the test. The distortions were applied to 10 HDR images in 100 distortion levels resulting a database of 6000 HDR test images. The subjects were asked to find the image JND location on each set of 100 images they had the freedom to explore. The effect of content features on the noticeable threshold selection is investigated per distortion type. Our results in some cases show a significant correlation between content features and JNDs. We are hoping that our results can contribute to further exploitation of a precise HVS model for HDR quality assessment and optimization of the coding and bit allocation in HDR compression.**

I. Introduction

Capturing natural scenes often involves situations where extremely bright and dark areas are meant to be recorded together. Regular imagery though has limited ability to record such a wide range of lighting levels. Often this results into loss of saliency and failing to capture the fine details in the low light areas. High Dynamic Range (HDR) imaging was particularly introduced to tackle this issue. An extensive amount of research has been devoted to the capture, process and display of the HDR content. This trend recently has increased the demand for automated visual quality assessment of HDR data by means of objective image quality measures (IQM). Several popular full reference IQMs like PSNR and SSIM can be adapted for HDR data, especially for the purpose of signal error analysis. More recently, methods like HDR-VDP-2 [1] have been introduced to exploit the differences between multiple luminance conditions (from photopic to scotopic vision) utilizing a more comprehensive model of the early stages of the HVS. Similar to the Low Dynamic Range (LDR) case, subjective tests are required to provide the ground truth for benchmarking and improving the accuracy of such IQM methods. For example in [2], a Double Stimulus Impairment Scale (DSIS) methodology was utilized to calculate the Mean Opinion Scores (MOS) of 50 HDR images. Those images were first converted into a set of LDR images by means of a tone mapping function and then compressed with JPEG, JPEG 2000 and JPEG XT encoders. Next, PSNR, SSIM and HDR-VDP-2 were used to predict the perceived quality and compared with the reference MOS. The PSNR and SSIM measures were adapted for HDR imagary using a logarithmic mapping as well as a Perceptually Uniform encoding [3]. They concluded that for the specific case of HDR backward-compatible compression, where the compression artifacts are very similar to the LDR case, the performance of the adapted versions of PSNR and SSIM can be as good as HDR-VDP-2 or even better while benefiting from significantly less complexity. Similar subjective tests were also conducted specifically for the case of compression artefacts inflicted by the JPEG XT encoder, which were reported in [4], [5] and [6]. Subjective tests were conducted to acquire the MOS scores for a set of HDR video clips in [7], [8] and more recently in [9].

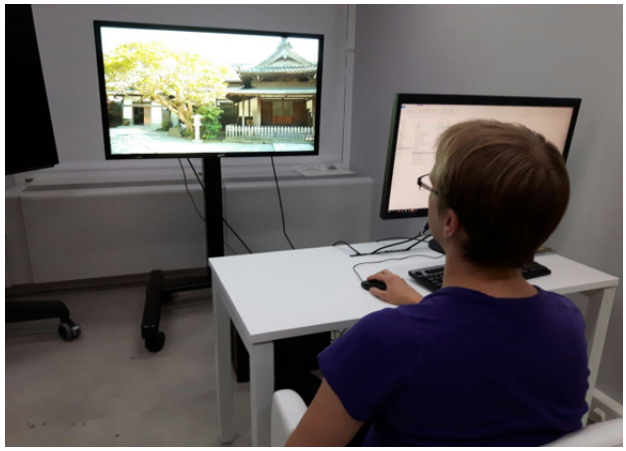


Fig. 1: The testing environment for the subjective test.

In this research, we investigate the HVS sensitivity to a set of common distortion types for HDR images with the hope to provide a better understanding of HVS behaviour in response to the HDR image degradations. We report the results of an extensive JND subjective test which was conducted for a selected set of HDR images. To the best of our knowledge, this is the first time that such a comprehensive JND test for HDR images in a systematic way is being conducted. Six distortion types including quantization errors, compression errors (JPEG and JPEG2000) and three processing distortions(Gaussian Noise, Blur and Contrast change) where applied on 10 HDR images

978-1-5386-2605-4/18/$31.00 ©2018 IEEE

in 100 levels of severity, resulting in a pool of 6000 HDR test images. The results of this experiment are expected to help in:

- Building more accurate HVS models tailored to HDR content for predicting perceptual quality;
- Optimizing the performance of HDR image encoders.

In the next section, we will detail the conducted subjective test. In section III, we provide the experimental results and analyze the outcomes. Finally, In section IV we provide concluding remarks.

II. Subjective Test Methodology

This section elaborates the JND-based subjective quality assessment methodology of HDR images.

A. Test environment

The subjective experiment was performed at the Visual Quality Test Laboratory of the Vrije Universiteit Brussel-imec, which complies with the ITU-R recommendations [10] for subjective quality assessment of visual content. A controlled lighting system with 6500K colour temperature is installed in the test room and only indirect lighting of the test screen is permitted.

A 47 inches SIM2 HDR47 screen, with viewing angle ± 85 degree (V/H), square pixel pitch of 0.5415 mm, 1920x1080 resolution and contrast ratio higher than 4×10^6 is utilized to display the test images to the subjects. The screen has back-light LEDs with 12 bits of brightness resolution for each LED, yielding a maximum luminance of 6000 cd/m2. The distance of the subjects from the screen is 3.2 times the picture height according to ITU-R BT.2022 standards[11].

A software tool was developed to display and record the JND threshold decided by subjects. The subjective test tool displays all the test sequences in a random order. For each source image and distortion type, the subject will scroll through 100 images arranged in quality descending order. The user has freedom to scroll back and forth between sequences of images to select the image of which the distortion starts to appear and to be noticeable. The selection can be performed simply by pressing the left mouse-button followed by pressing the right mouse-button. Fig 1 shows the test room.

B. Stimuli

Ten high-quality HDR images of various dynamic ranges were used in the experiment. The images are chosen from two public HDR datasets, namely the HDR-Eye [12] and Fairchilds HDR Photographic Survey datasets[13]. In particular, six images from HDR-Eye (1920x1080 resolution) and four images of the Fairchild dataset (larger than 4K resolution) were used in the experiment. The images of Fairchild were downscaled to meet the Full-HD size of the screen. Fig.2, displays the selected images for the JND-based subjective experiment. The images are tone-mapped for illustration purposes. The dataset includes scenes with different illumination patterns (daylight, highlights and shadow, night scenery).

The JND test was conducted on images of different distortion types to analyze the HVS sensitivity to various distortion

Fig. 2: The 10 HDR images used in the test were resized to 1920x1080 resolution.

artefacts in the context of HDR. The Full-HD images are subjected to six types of distortion: JPEG, JPEG 2000, Contrast Change (CC), Gaussian Blur (GB), Gaussian White Noise (GWN), and quantization error. From each source image, 100 images are generated in which the distortion level is increasing from level 1 to 100. In total 6000 images were generated(10 source images x 6 distortion types x 100 severity levels).

- The **JPEG XT** coding scheme (ISO/IEC 18477) is a new standard of the JPEG committee, which offers backward compatible compression of HDR content and which has been utilized to generate JPEG distorted HDR images. JPEG XT performs coding in base and enhancement layers. The reference coding scheme that we utilized for the test[1] uses two parameters QFb and QFe (ranging from 1 to 100) to control the quality of base and enhancement layers, respectively. Here, we encode the source

[1]JPEG document WG1N6639 in the JPEG document repository, version 0.8 (February2014).

TABLE I: The minimum and maximum PSNR values per image and distortion types, representing the related severity range of the applied distortions.

	Quantiztion		Blur		contrast		JP2k		JPEG		Noise	
	min	max	min	max	min	max	min	max	min	max	min	max
Car507	51.44	82.78	42.37	88.85	32.10	62.68	41.3	83.29	22.68	58.05	52.8	91.5
BigfootPass	51.31	76.13	34.67	80.35	29.32	56.91	33.33	73.16	15.23	50.38	51.81	83.53
C01_HDR	53.03	83.98	47.09	84.81	31.85	62.3	48.77	83.94	11.02	57.54	53.08	89.6
C11_HDR	51.68	82.41	35.17	85.2	32.10	61.64	34.22	80.47	19.54	52.44	52.57	88.17
C12_HDR	50.98	83.08	34.44	90.65	36.29	66.53	34.39	83.55	23.44	56.83	53.13	92.58
C17_HDR	50.88	85.5	41.63	88.82	34.02	64.07	44.1	93.96	25.07	62.44	52.88	93.12
C24_HDR	51.74	83.53	33.87	85.2	33.58	63.29	33.44	83.54	22.21	55.28	52.46	94.34
C30_HDR	51.69	80.35	36.77	84.35	31.45	62.69	37.45	81.36	16.08	57	52.46	90.05
PaulBunyan	51.74	80.26	43.13	87.8	31.65	62.56	39.99	82.74	26	58.79	52.52	91.05
PeckLake	51.6	80.48	36.71	83.82	29.44	59.32	33.72	80.34	21.14	54.36	52.39	88.05

images with $(QFb, QFe) = (1,1),(2,2),...(100,100)$ to produce 100 JPEG distortion levels. The JPEG distortion mainly induces blocking artefacts to the image.

- **JPEG2000** compression is applied using the HDR Toolbox of Matlab v 1.1.0 [14]. The *HDRJPEG2000Enc* function of the toolbox was used to encode source images in which the compression ratio is adjusted to produce images with 100 quality levels. Finally, the *HDRJPEG2000Dec* function is used for decoding. The JP2K compression distortion typically appears as blur and ringing artefacts in images.
- **Noise**: to add noise to source images, a standard normal pdf of variance σ_N^2 is applied to RGB planes using the *imnoise* command in Matlab. The same variance was applied to the R, G, and B components. The variance is altered on a logarithmic scale to cover a sufficient range of quality levels and to provide gradual quality changes.The variance is changed in logarithmic scale from 5×10^{-13} to 5.6×10^{-6} to cover a sufficient range of quality levels and provide gradual quality changes.
- **Blur**: the blurred images were produced by employing a circular-symmetric 2-D Gaussian kernel.The Matlab *fspecial* and *imfilter* commands were used and R, G, and B channels were exposed to the same kernel.The standard deviation is increased from 0.1 to 5 to generate blurry images.
- **Contrast**:The Matlab *localcontrast* command was utilized to apply contrast changes to input images where the enhancement parameter is linearly changed in range of [0.0001 0.005] to degrade the contrast of the source images.
- **Quantization Error**: Finally, banding artefacts were applied by modifying the quantization values. The Matlab *imquantize* command is used to reduce the number of discrete levels and generate the artefact. The quantization level is decreased logarithmicaly in range of 20000 to 220.

As a measure of the overall energy degradation applied by each of the distortions, in Table I, we have provided the PSNR values for the images with lowest and highest amount of each distortion. It should be noted that, as it is explained above, all the images have gone through the same process for imposing the degradation. The variation in their PSNR values though depends on their content, yielding different PSNR ranges.

TABLE II: Overall Brightness (OB), Dynamic Range (DR), Contrast Parameter (CP) and Spatial Complexity (SC) of the raw HDR images. The images in the table are sorted based on their OB values. The smallest and largest values per parameter are boldfaced.

Image Name	OB	DR	CP	SC
C24_HDR	**0.282**	2.045	0.376	0.025
C12_HDR	0.468	3.794	0.752	**0.093**
C11_HDR	0.533	2.550	0.672	0.090
C30_HDR	0.533	2.719	0.630	0.018
PeckLake	0.547	2.555	0.640	0.037
C17_HDR	0.548	3.439	0.793	**0.013**
Car507	0.548	2.977	0.701	0.058
BigfootPass	0.583	**0.974**	**0.259**	0.077
PaulBunyan	0.659	2.011	0.480	0.029
C01_HDR	**0.685**	**12.741**	**5.179**	0.057

C. Observers and scoring protocol

A total number of 18 subjects among which 11 male and 7 female participated in the test. The subjects age ranged between 22 and 36 years and the average age was 28.6. All subjects passed the visual acuity and colour vision tests through Snellen and Ishihara charts. A training session was organized before the test to educate the subjects on the test procedure and subjective test tool. We displayed 60 test sequences randomly. Each sequence included 100 images in which the quality was monotonically decreasing. It is expected that humans are not able to differentiate all 100 levels. The subjects were asked to browse through the images and find the first one that exhibits a just noticeable difference. In other words, the subjects identified the threshold (JND) point in the image sequence were the artifacts are starting to become visible. The test duration for each subject was about 15 to 25 minutes depending on how fast they choose the JND locations in each sequence of images.

D. Quantitative content features

To ensure the diversity of the stimuli, we mainly followed the procedure introduced in [2] to select the images. A series of content features was considered to guarantee the diversity of

the stimuli. These features were also utilized in experimental analysis. They included:

1) **Overall Brightness(OB)**, which is quantified by:

$$\mathbf{OB} = \frac{\log L_{avg} - \log L_{min}}{\log L_{max} - \log L_{min}} \tag{1}$$

where the average luminance is computed as $\log L_{avg} = \sum_{ij} \log\left(L(i,j) + \delta\right)/N$, N is the image pixel count and $L(i,j)$ the luminance of pixel (i, j). A small constant δ is added to deal with the singularity case of the black pixels. Minimum and maximum luminance is depicted as L_{min} and L_{max}. To avoid outlier pixel values, the brightest and darkest percentile of image pixels were excluded from the calculations.

2) **Dynamic Range(DR)**: The dynamic luminance range of each image in logarithmic scale calculated as:

$$\mathbf{DR} = \log \frac{L_{max}}{L_{min}} \tag{2}$$

3) **Spatial Complexity(SC)** was calculated as:

$$\mathbf{SC} = \sigma^2(Sobel(I_{maped})) \tag{3}$$

where the standard deviation (σ^2) of Sobel operator was calculated for each image after tone-mapping them into LDR image shown here as (I_{maped}).

4) **Contrast Parameter(CP)**: An additional parameter was defined to represent the overall contrast of each image:

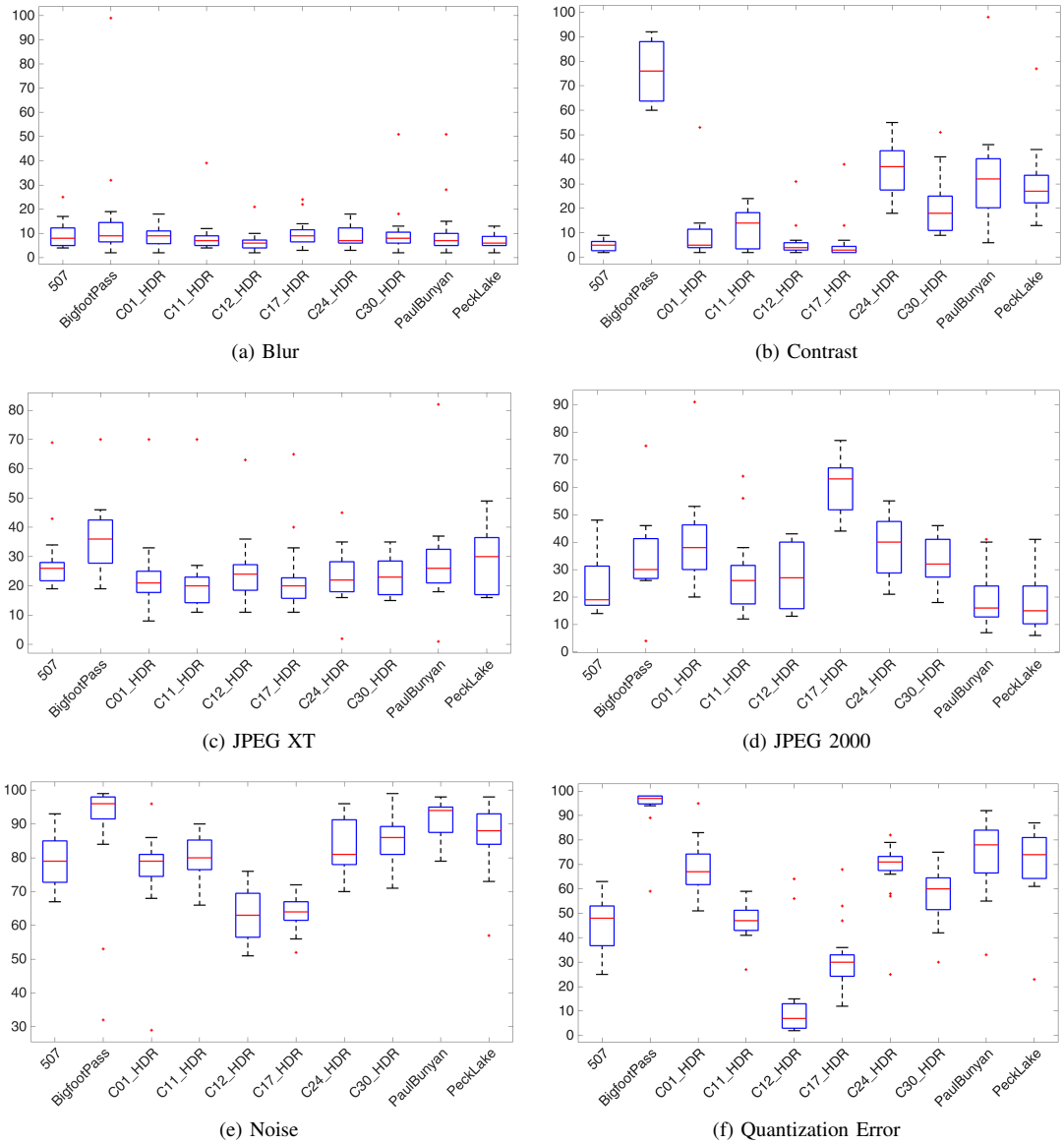
$$\mathbf{CP} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log L_i - \log L_{avg})^2} \tag{4}$$

Table II shows the values of above parameters for the reference HDR images, i.e. before adding the distortions.

III. Results and Analysis

After the data gathering step, we generated the box plots of all JND values per distortion type given to each image sample, shown in Fig 3. For Blur shown in Fig 3.a, the box sizes are very small and their whiskers are also very short which demonstrates that all of the human subjects had a close agreement on finding the JND points for all images. The median of the JNDs for all images(red lines in the box plots) are also very close to each others. This may imply a significant level of content independency for detection of blurriness.To a lesser degree, the same behaviour can be seen for JPEG XT, depicted in Fig 3.c where also more variance between subjects' opinions can be seen(larger interquartile ranges and whiskers). High variance in case of Contrast (Fig 3.b) though seems more regulated where the boxplots can be clearly categorized into two groups with BigfootPass being an outlier distinctively. Looking at their parameter values, BigfootPass has the lowest DR and CP among other test images and relatively high SC. On the other hand, images C01, C11,C12, C17 and Car507, with lowest JND have significantly high DR, CP and lowest SC values. An interpretation of this situation can be that for images with overall lower dynamic range and covered with highly textured areas and minimalistic smooth areas(e.g. BigfootPass) strong contrast masking is present and significantly increase the noticing threshold when the contrast is decreased. The other side is when dynamic range is very high (e.g.C01HDR) and/or large smooth areas with less spatial complexity are available (e.g. Car507). Content dependence is also present for JPEG2000, where detection of the noticeable thresholds were more troublesome for the subjects(refereing to large box plots for all images). A similar situation is observable for the case of Noise. Looking at the Fig 3.f, it immediately clarifies that JND threshold is highly dependent to the content of the displayed image where completely different JND levels have been selected by subjects per image. Also the interquartile ranges vary from image to image, which shows different degree of agreement on the detected JND thresholds.

To further elaborate on the effect of the image characteristics on the JND levels, we attempted to quantitatively identify the relation between the selected JND levels and the four content features introduced in section II.D. To do this, we calculated the Spearman Rank Order Correlation (SROCC) between the set of all JND values per distortion type and each one of the four content features. To stabilize the outcomes we performed an outlier removal step on the raw JND scores such that for each image individual JND scores higher or lower than the first decile of the JNDs for that image where removed from the data. Next, we calculated the mean, median and standard deviation(STD) of the JNDs for each image and calculated the SROCC between each one of them and each of the four content features. The results are shown in Fig 4. It is interesting to see that for the case of Noise, Quantization Error, Contrast change and to a lesser degree for JPEG XT, the DR and CP features show a remarkably high negative correlation w.r.t. both mean and median JND ranking orders. This implies that in a set of HDR images contaminated with the mentioned distortion types, an ordering based on their DR and/or CP values in descending order, their ranks are highly similar to the case of ordering them based on their JND threshold point. However, DR and CP seems to have a very limited to nonexistent correlation for the case of images distorted by Blur and JPEG 2000 (absolute correlations lower than 0.5 are close to comparing JNDs with random ranking for the limited size of our data pool and thus shall not be utilized to draw any conclusion). Instead, for the case of Blurriness, the OB depicts a high correlation with the JND ranking order, and is thus appropriate to predict the JND level of blurred images. The results shown in Fig 4.c are though more difficult to interpret. Basically, the STD of JNDs shows the level of certainty and agreement between the subjects in choosing the noticeable distortion level.(higher STD value in this case shows higher diversity and uncertainty in the opinions.) The SROCC in this case quantifies the correlation between ranking images according to STD(uncertainty level of opinions) and the content features. For example for the case of contrast change, where a strong inverse correlation exists, one may conclude that by ordering the images in a descending order based on the DR and/or CP values, their ranks are highly



(a) Blur (b) Contrast

(c) JPEG XT (d) JPEG 2000

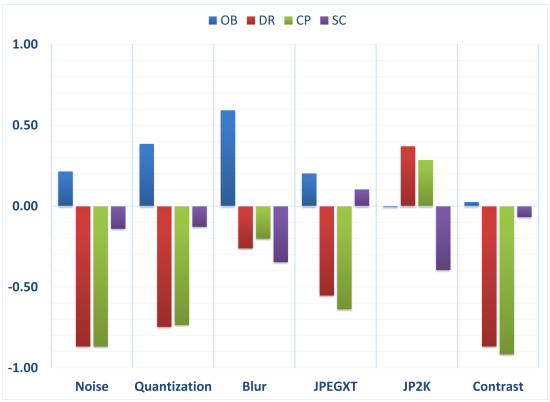
(e) Noise (f) Quantization Error

Fig. 3: Boxplots of the selected JND points (shown in Y axis) in the range of 1 to 100 per image sample. Each boxplot shows the values for one of the distortion types.

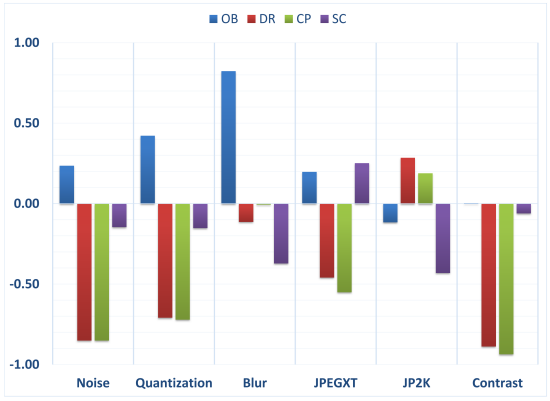
similar to the case where we order them based on the variance of JND values. A final remark regarding the results shown in Fig 4 is related to the rather low overall correlation between the SC features and JND statistics. According to these results, it seems that the current representation for the spatial complexity is barely sensitive to the variations of JND for HDR images. This either implies that, spatial complexity has a minimalistic role at defining the noticeable threshold of tested distortions for HDR images, or may come from the fact that the utilized SC feature poorly represents the actual spatial complexity of the scene at least for the case of HDR imaging. Further investigations are required to clarify this ambiguity.

IV. Conclusion

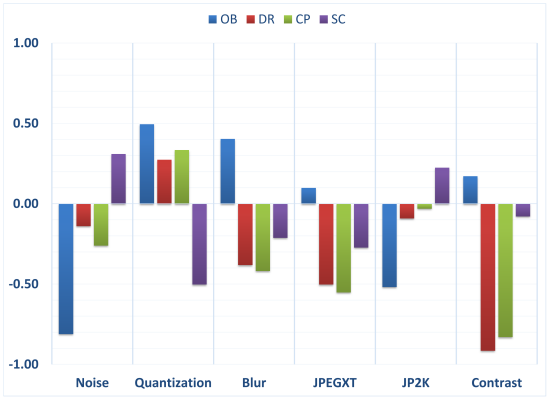
A subjective JND test was performed on a set of HDR images distorted with six different distortion types namely noise, quantization, JPEG XT, JPEG 2000, contrast change and blur. The distortions were applied using 100 levels of severity and the subjects were asked to find the noticeable threshold by freely scrolling through the 100 instances. The analysis of the results demonstrated that distortions like blurriness are detected with high certainty by the subjects irrespective of the image contents. On the other hand, the detection threshold for contrast changes and quantization appeared to be highly dependent on the characteristics and content of the HDR images. In an attempt to quantify those image characteristics, we defined four content features and examined their correlation with the JND mean, median and standard deviation. The results illustrate that content features like dynamic range and contrast have a strong inverse rank order correlation with respect to the JND mean and median for most of the distortion types. The overall brightness feature also depicted a high



(a) Mean JND



(b) Median JND



(c) STD JND

Fig. 4: SROCC calculated between (a)mean, (b)median and (c) standard deviation of JNDs per distortion type versus 4 content features including Overall Brightness (OB), Dynamic Range (DR), Contrast Parameter (CP) and Spatial Complexity (SC) of the raw HDR images. Negative SROCC represents inverse correlation while higher absolute value of SROCC represents better compatibility between ranking order presented by given JND points and the values of the specified content feature.

correlation w.r.t the JND threshold for the case of blurriness. The spatial complexity feature did not result in any significant correlation and hence appeared to be irrelevant in the current experimental setting. Also, box plots of the raw test results per distortion type have been provided to facilitate the means for the readers to draw further conclusions of their owns. We are expecting that these results can be utilized both for optimizing the performance of the HDR encoders as well as helping to design efficient perceptual quality predictors for HDR contents.

Acknowledgements

The research leading to these results has received funding from Horizon 2020 Research and Innovation Programme under Grant Agreement N688619 (ImmersiaTV) and imec-icon HD2R project.

References

[1] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4. ACM, 2011, p. 40.

[2] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for hdr image compression," 2014.

[3] T. O. Aydin, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full dynamic range images," in *SPIE Human Vision and Electronic Imaging XIII*, 2008.

[4] P. Hanhart, M. V. Bernardo, P. Korshunov, M. Pereira, A. M. Pinheiro, and T. Ebrahimi, "Hdr image compression: a new challenge for objective quality metrics," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 159–164.

[5] C. Mantel, S. C. Ferchiu, and S. Forchhammer, "Comparing subjective and objective quality assessment of hdr images compressed with jpeg-xt," in *Multimedia Signal Processing (MMSP), 2014 IEEE 16th International Workshop on*. IEEE, 2014, pp. 1–6.

[6] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of hdr images compressed with jpeg xt," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[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," in *International Conference on Multimedia Signal Processing (ICMSP)*, 2014.

[8] K. Minoo, Z. Gu, D. Baylon, and A. Luthra, "On metrics for objective and subjective evaluation of high dynamic range video," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2015, pp. 95 990F–95 990F.

[9] R. Mukherjee, K. Debattista, T. Bashford-Rogers, P. Vangorp, R. Mantiuk, M. Bessa, B. Waterfield, and A. Chalmers, "Objective and subjective evaluation of high dynamic range video compression," *Signal Processing: Image Communication*, vol. 47, pp. 426–437, 2016.

[10] B. Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 13–500, 2012.

[11] ITUR-BT2022, "General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays," *International Telecommunication Union*, 2012.

[12] H. Nemoto, P. Korshunov, P. Hanhart, and T. Ebrahimi, "Visual attention in ldr and hdr images," in *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, no. EPFL-CONF-203873, 2015.

[13] M. D. Fairchild, "The hdr photographic survey," in *Color and Imaging Conference*, vol. 2007, no. 1. Society for Imaging Science and Technology, 2007, pp. 233–238.

[14] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, "Advanced high dynamic range imaging: Theory and practice. isbn: 978-156881-719-4, ak peters."